Kathleen B. Levitz
Vice President-Federal Regulatory

BELLSOUTH

Suite 900
1133-21st Street, N.W.
Washington, D.C. 20036-3351
202 463-4113
Fax: 202 463-4198
Internet: levitz.kathleen@bsc.bls.com

EX PARTE OR LATE FILED

December 2, 1998

RECEIVED

DEC 2 - 1998

FEDERAL COMMUNICATIONS COMMISSION
OFFICE OF THE SECRETARY

Ms. Magalie Roman Salas
Secretary
Federal Communications Commission
1919 M Street, NW, Room 222
Washington, D.C. 20554

Re: Written Ex Parte in CC Docket No. 98-56 and CC Docket No. 98-121

Dear Ms. Salas:

On November 30, 1998, BellSouth filed with you notice of a written ex parte it had made that day to Carol Mattey, Chief of the Policy and Program Planning Division. That ex parte consisted of a copy of the Interim Statistical Analysis for BellSouth Telecommunications, Inc.(BellSouth filing), submitted to the Louisiana Public Service Commission on November 19, 1998, in that Commission's Docket No. U-22252-Subdocket C. We subsequently learned that the appendices in the copy of the BellSouth filing sent to Ms. Mattey and to you were incomplete. Today we sent Ms. Mattey a complete copy of the BellSouth filing. We are now filing with you this notice of that second written ex parte.

Pursuant to Section 1.1206(a)(1) of the Commission's rules, we are filing two copies of this notice and the corrected copies of the BellSouth filing for both of the dockets identified above. Please associate this notification with the record in both those proceedings.

Sincerely,

Kathleen B. Levitz

Kathleen B. Levitz
Vice President-Federal Regulatory

Attachment

cc: Carol Mattey (w/o attachment)

BELLSOUTH

Kathleen B. Levitz
Vice President-Federal Regulatory

Suite 900
1133-21st Street, N.W.
Washington, D.C. 20036-3351
202 463-4113
Fax: 202 463-4198
Internet: levitz.kathleen@bsc.bls.com

December 2, 1998

Ms. Carol Mattey, Chief
Policy and Program Planning Division
Common Carrier Bureau
Federal Communications Commission
1919 M Street, NW, Room 222
Washington, D.C. 20554

Re: Written Ex Parte in CC Docket No. 98-56 and CC Docket No. 98-121

Dear Ms.Mattey:

On November 30, 1998 BellSouth sent to you a copy of a recent filing made by BellSouth in the Louisiana Public Service Commission's Docket No. U-22252-Subdocket C. The document, entitled "Interim Statistical Analysis for BellSouth Telecommunications, Inc.," had been submitted to the Louisiana Public Service Commission on November 19, 1998. We subsequently learned that the appendices in the copy sent to you were incomplete. Attached is a copy of the BellSouth Louisiana filing that contains copies of all the appendices included in that filing.

Pursuant to Section 1.1206(a)(1) of the Commission's rules, I am also filing with the Secretary the required notices and copies of this written ex parte presentation in both the dockets identified above.

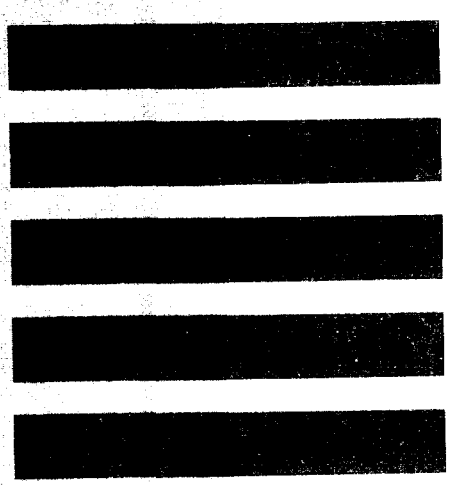
Sincerely,

Kathleen B. Levitz

Kathleen B. Levitz
Vice President-Federal Regulatory

Attachment

cc: Andrea Kearney Jake Jennings Michael Pryor



Interim Statistical Analysis

For

BellSouth

Telecommunications, Inc.

by

Susan Hinkins, Edward Mulrow and Fritz Scheuren

Submitted to:

Louisiana Public Service Commission (LPSC), Docket U-22252 Subdocket C

Submitted on:

November 19, 1998

ERNST & YOUNG LLP

TABLE OF CONTENTS

Appendices

List of Tables and Figures

Provisioning and Order Completion Interval

Maintenance Average Duration

OSS Average Response Interval

Geographical Disaggregation

Interim Conclusion

Executive Summary

The Louisiana Public Service (LPSC), under Docket No. U-22252, requested BellSouth to set out its views on the "application of a statistical analysis to performance measurement data." Our findings are summarized under three broad headings:

(1) Pros and cons on alternative testing methodologies,

(2) Actual results from these tests for three performance measures, and finally,

(3) The levels of disaggregation and testing that would be statistically valid.

In what follows, we quote the LPSC charge in italics under each of these.

Alternative Testing Methodologies

Parties should be prepared to discuss pros and cons of each statistical testing methodology: modified z [the LCUG test], pooled variance [the FCC test], and ... [BST alternatives.]

For the most part we found that the LCUG and FCC tests had to make strong assumptions that did not appear warranted in the data we examined. The BST tests, on the other hand, are not subject to such strong assumptions.

In summary, the BST approaches appear to work efficiently and can be interpreted as a safe starting point for statistically analyzing differences, if any, between CLEC resale and BST retail customers. This is simply not the case for the LCUG and FCC calculations. Table 1, which follows, provides this summary in tabular format, addressing the specific dimensions quoted as the outset of this report from Louisiana Docket No. U-22252.

Actual Test Results

BellSouth to report on the results of different statistical tests ordered by the Commission.

The three tests, FCC, LCUG, and BST, were examined for their performance on two data sets, provisioning – order completion interval (OCI), and maintenance – maintenance average duration (MAD). In the case of the Average Operating Support Services (OSS) Response Interval, we had only daily summary averages for BST and for the CLECs. Because of this, the LCUG and FCC tests could not even be calculated. We were, however, able to successfully make use of time series techniques to analyze the data.

Table 2 provides a tabular summary of our analysis. For two of the three measures, the weight of the evidence gives no strong sign that something other than full parity (or better) exists between BST retail and CLEC resale customers. These measures are Maintenance Average Duration and Average OSS Response Interval:

- For maintenance average duration, the evidence indicates a slight but arguably not statistically significant difference in "favor" of BellSouth for August and a slight but again arguably not a statistically significant difference in "favor" of the CLECs in September.

- For the OSS response interval, the evidence is quite strong that, for August at least, BellSouth could be favoring the CLECs over its own customers.

For the third measure -- order completion interval -- the evidence supports a different conclusion. As the report details, BellSouth appears to be providing service to the CLECs which is statistically significantly slower than it provides to its own retail customers. This difference is not large overall for August. After adjusting for customer mix the difference turns out to be just 0.15 days in August. However, this difference rises to 0.59 days in September, clearly both statistically and operationally significant (and thus warranting further study of the underlying causal structure).

Levels of Disaggregation

Parities should be prepared to discuss what levels of disaggregation would make statistically valid sample.

From a geographical standpoint, we believe that test statistics should be reported at only the state and LATA levels. It has been suggested that Metropolitan Statistical Areas (MSAs) also be used as a reporting level below the state level. However, we believe that it makes more sense to use LATA. Our reasoning was as follows:

- LATAs are a meaningful geographic business unit for BellSouth. MSAs are statistical entities, subject to redefinition by the Office of Management and Budget (OMB) of the Federal government. In fact, there will be a major revision of MSAs in connection with the 2000 Census. Hence MSA units may not have a stable definition over time.

- MSA's in Louisiana vary considerably in size as measured by the number of wire centers servicing them. This has the effect for several of making the sample sizes available "small" – too small to safely employ the types of statistical test discussed in this report.

When it comes to making judgements as to whether or not BellSouth is meeting its nondiscriminatory obligation with respect to the service it provides CLECs and their customers, there are potential problems that can arise when the results of too many parity tests are aggregated. These problems include: dependencies that exist between parity tests, dependencies between consecutive monthly measurements, and parity measures with non-normal distributions.

Our analysis indicates that these problems are negligible when the results of only five to ten parity tests are aggregated in any given month. Multiple tests above this level raise very serious theoretical, as well as, management issues.

More to Come

This report is of work in progress. For example, by the time of the November workshop we plan to move from our present interim stage to a point where we will make recommendations to the Commission. At that time, we will also present additional analyses based on what we have learned regarding the statistically significant differences found for the order completion interval and its potential causes and implications.

Table 1. – Interim Summary of Required Methods Comparison Made for the Louisiana Commission under Docket U-22252

Testing Proposal	When ILEC & CLEC processes are different and not expected to yield same results	When ILEC actually is employing discriminatory practices.	When assumptions necessary for the statistical test to be valid are not meet
LCUG	Calculating these measures at the level of descriptive reporting required can lead to comparisons that are not "like-to-like." The answer here is not more detail (which pushes against sample size limits) but an analytic summary based on standardized data. That is the approach we have taken. In particular by building upon the CLEC volumes to standardize the BST comparisons, much of this concern can be reduced or avoided.	This test has possible merit and in some settings might even be preferred to that suggested by the FCC, albeit the FCC and LCUG numerical results we saw are virtually identical in most cases and have about the same problems -- notably that the strong assumptions required for validity do not always hold.	For monthly Louisiana results clear evidence exists that the assumptions in the LCUG test fail to hold and, hence this test is invalid for general use. Moreover it cannot be employed at all to statistically study differences in OSS response intervals between BST and the CLECs.
FCC		This measure could work well, if "likes-to-likes" are compared. Required, though, is that strong assumptions hold for it to be valid – something we did not find always to be the case.	This test has the same basic weaknesses as the LCUG approach and is, hence, also unsuitable for general use. Moreover, it makes an additional assumption which does not appear to hold in all settings.
BST		The methods we have recommended will have essentially the same efficiency (or power) as the FCC and LCUG tests to detect differences, should they exist. They are, moreover, completely practical and do not prefer one side over the other.	For individual Louisiana results, possible assumption failures are judged unlikely and no evidence for them was found. For the month-to-month changes more study is needed and this will be covered at the November 30 workshop.

Table 2. – Summary Results of Preferred Testing Approach by Type of Performance Measurement, August and September Separately

Performance Measurement	Difference of "Likes-to-Likes"	BST Test Statistic	Interpretation
Order Completion Interval - Provisioning			For both August and September, the tests done show that statistically significant differences exist favoring BellSouth over the CLECs. For September, moreover, the difference almost certainly are large enough to have operational significance. Both months merit further study and our findings will be given at the November 30th workshop.
August	-0.14 Days	-2.57	
September	-0.59 Days	-8.81	
Maintenance Average Duration	-1.38 Days 2.32 Days		The test statistics for the Maintenance Average Duration are near statistical significance in each month but in opposite directions. No further action seems called for.
August		-1.93	
September		2.43	
OSS Response Time	.3197 Seconds .1028 Seconds		For OSS Response Time, the test statistics are both positive and for August highly significant, suggesting if anything, that BellSouth is favoring the CLECs over itself.
August		3.78	
September		1.20	

Note: "Statistical Significance" in this report is defined to have been reached when the test statistic is outside the range ± 2. By convention, when the difference is positive, we say the measure suggests that the CLECs resale customers are getting better treatment than BST retail customers. The reverse is true if the sign of the difference is negative. Differences that are +2 or larger are defined therefore to be differences which statistically significantly "favor" the CLECs. Differences that are -2 or smaller are defined to be differences which statistically "favor" BellSouth (see Glossary and Appendix B).

Interim Statistical Analysis For BellSouth Telecommunications, Inc. Submitted to Louisiana Public Service Commission (LPSC) Docket U-22252 Subdocket C

I. Introduction and Scope

BellSouth has been asked by the Louisiana Public Service Commission (LPSC), in Docket No. U-22252 - Subdocket C (dated August 12, 1998), to set out its views on "the application of a statistical analysis to performance measurement data" (Ibid., page 15). The present report is intended to provide an interim response to that request. We will also address the wider context within which such performance measurements might be, to again quote the order, "useful in determining whether BellSouth is meeting the statutory requirements with respect to its provision of unbundled network elements, resale, and interconnection to CLECs" (Ibid., page 15).

The setting for the analysis is crucial to the interpretation of any statistical significance that might be found. There is no doubt that "statistical analysis can help reveal the likelihood that reported differences in an ILECs performance toward its retail customers and CLECs are due to underlying differences in behavior rather than random chance" (Ibid., pages 15 - 16).

The Louisiana Public Service Commission under Docket No. U-22252 states "that a uniform methodology which identifies those items which need to be measured, how they are to be measured, and how the results are to be reported is also desirable and would be beneficial to all parties" (Ibid., page 16). We agree with this goal as well, stipulating only that the use of a single method may not be desirable while a single methodology (or a set of methods) could be. In particular, we propose a family of methods in implementing the statistical analyses we will be presenting. To frame our presentation the next paragraph from the LPSC Docket U-22252 is quoted in its entirety.

> **"Statistical tests are effective in identifying those measurements where differences in performance exist. The tests themselves cannot identify the cause of the apparent differences. The differences may be due to a variety of reasons, including: 1) when the ILEC and CLEC processes being measured are actually different and should not be expected to produce the same result, 2) when the ILEC is employing discriminatory practices, or 3) when assumptions necessary for the statistical test to be valid are not being met."** (Ibid., page 16)

Apparent statistically significant differences in BellSouth and CLEC performance can arise for any of these reasons. To meet the Louisiana Commission's purpose we will recommend techniques that are robust (or safe) in the presence of possible assumption failure, carefully examine BellSouth Telecommunications (BST) and CLEC performance so "like" is compared only to "like," and are still able, in a highly efficient manner, to detect differences. Those differences, if any, could be

connected to service performance differences. Upon investigation those differences could lead to concerns about possible discriminatory practices should these exist.

Along with the BellSouth approach, the Commission ordered that two other testing procedures be examined: the LCUG modified Z-test endorsed by the CLECs, and the pooled variance Z-test offered by the FCC in its **Notice of Proposed Rulemaking (Appendix B)**, "so the competence of each test can be demonstrated over a reasonable period of time." As further requested, BellSouth has obtained and analyzed "its proposed statistical test, the modified Z-test endorsed by LCUG, and the FCC's proposed pooled variance test, ... for the following performance measurements which compute an average: Average OSS Response Interval – PreOrder and Ordering, Average Completion Interval-Provisioning, and Maintenance Average Duration" (Ibid., page 17).

This report fully complies with the Commission's request for this three-way comparison. Even so, we have indicated above that the report is an interim one. Part of our reason for using the word "interim" is that the Study so far has been confined to Louisiana performance measures selected by the Commission just for August and September 1998. We expect by the time that the Commission meets on November 30 that we will also have results for October. Three months of data may be judged by the Commission as reasonable, but in our view, because of the evident seasonality and other factors, an even longer period of time might be needed and we would be prepared to come back again in February, if required, to present six months of data.

To be responsive to the Commission, we have divided our discussion into seven sections and eleven appendices. The contents of each of these are briefly mentioned below -- first for the main report and then for the extensive supporting appendix materials:

For the main report, this section (Section I) introduces our work and sets out the required scope. The next three sections (Sections II through IV) take up in turn each of the three measures that the Commission requested be subject to statistical analysis: Order Completion Interval-Provisioning (Section II), Maintenance Average Duration (Section III), and Average OSS Response Interval – PreOrder and Ordering (Section IV). Section V summarizes some of the results geographically within Louisiana using the four local access transport areas (LATA): Baton Rouge, Lafayette, New Orleans, and Shreveport. In Section VI we summarize, as requested by the Commission, our views on the degree of disaggregation that can be usefully analyzed statistically. The final section (Section VII) provides our interim conclusions, based on what we have learned so far.

The eleven appendices cover the credentials and experience of the senior Ernst and Young statisticians who developed this report (Appendix A), provide details on the statistical calculations and derivations for Order Completion Interval-Provisioning and Maintenance Average Duration (Appendix B), present an extensive set of detailed statistical displays for these two measures for both August and September (Appendices C through F), look at geographic data by LATA for August and September (Appendices H and I), present a time series analysis of the data we were provided on Average OSS Response Times (Appendix G), speak to the technical issues of disaggregation and multiple testing – including the presentation of the extensive simulations we ran

to illustrate our point of view (Appendix J), and, finally, provide a glossary of acronyms and statistical terms used (Appendix K).

II. Provisioning – Order Completion Interval (OCI)

Since March of this year, the Louisiana Public Service Commission has been provided with a detailed view of the Order Completion Interval (measured in days) at the aggregate CLEC and BellSouth level for the state as a whole. For example, the August Service Quality Measurement reports provided the information summarized below on orders divided by dispatch, non-dispatch, residential, business, less than 10 circuits (Ckts), and more than 10 circuits. Figure 1 is an extract from the August report focusing on the dispatched orders. The non-dispatch orders are also included in the overall measure but are not shown in Figure 1. The complete set of provisioning measures reports for August and September are provided in Appendices C and D.

Figure 1 - Extract of August Order Completion Interval Report for Dispatch Orders (Averages in Days for Orders Under Ten Circuits (Ckts) and Orders for Ten Circuits or More)

	< 10 Ckts	>= 10 Ckts
[redacted]		
Louisiana		
[redacted]		
Louisiana		
- Resale Residence	4.09	5.00
- Resale Business	3.81	5.83
- UNE Loops w/ LNP		
[redacted]		
Louisiana		
- Retail Residence	5.38	4.69
- Retail Business	7.37	15.29

	< 10 Ckts	>= 10 Ckts
[redacted]		
Louisiana		
[redacted]		
Louisiana		
- Resale Design	19.70	23.33
- UNE Design	10.86	2.00
- UNE Non-design	10.18	0.00
[redacted]		
Louisiana		
- Retail Design	23.00	31.80

The information in these individual summaries is difficult to interpret in the context of determining if statistically significant differences in performance exist for similar customers. For one thing, it is desirable to have an overall view of the data so that a combined comparison can be made. This is not as easy as it sounds. In this section, five statistical tools are provided to help examine performance differences between aggregate CLECs and BST customers on provisioning and to help interpret them.

The five tools are a combination of statistics and graphics. The first four are primarily descriptive and interpretive tools, with only the last being a series of formal tests of statistical significance. The four descriptive measures are the mean for the order completion intervals; their corresponding standard deviations; a graphical presentation of the relative frequency distribution of the completion interval data; and, to help the eye see differences between the CLECs and BellSouth distributions, a quantile comparison of CLECs and BellSouth. The final tool is a set of three different statistical tests, one proposed by the LCUG, one by the FCC, and one by BellSouth.

Overall CLEC and BellSouth Means. – The arithmetic mean is a well-known descriptive measure for a set of observations. Differences in the CLECs and BellSouth mean order completion interval are a natural place to start when looking for possible differences in treatment. It would be reasonable if there were no service differences to expect that, on average, resale orders from CLECs take about the same amount of time to complete as do retail orders at BellSouth. Figure 2, below, shows the overall means of the order completion interval for BellSouth and the CLECs for the month of August.

Figure 2 - Order Completion Interval
Difference in Means (August)

Service Provider	Mean (Days)
BST	1.20
CLEC	1.62
Difference	-0.42

At first glance, there does seem to be a noticeable difference between the amount of time required to complete an order for the CLECs and the amount of time required for BellSouth. Here, however, a true difficulty arises in taking the data at their face value. BellSouth is a company that operates throughout the entire state. By contrast, an individual CLEC may be confined to operating in a relatively small region, and the types of services it offers may be limited as well. What this implies is that we could be comparing very dissimilar entities, with very dissimilar volumes of business. A storm in one parish can have a significant effect on the overall CLEC measure, while that same effect on BellSouth, although present, is obscured due to the large volume of business outside that area that acts to overshadow it.

This points to a need to eliminate as much underlying dissimilarity as possible before making comparisons. This much is intuitively clear, and it is actually already in practice, in the form of the monthly SQM reports which break down each measure to some combination of different levels. However, additional adjustments need to be made to simply account for the difference in volume of business between BellSouth and the CLECs at each of the specific levels. To explain this point, we have set out an illustration in the box below.

Illustration A—Method of Standardizing to Compares Likes-to-Likes

Statistical adjustment attempts to account for interpretation issues arising from different relative volumes. Suppose we have just two groups: new orders, and change orders. Suppose further that new orders always take 2 days to complete, and change orders always take 1 day, no matter what the source (Provider A or Provider B).

The numbers of new and change orders for each provider are given in the table below.

Service Provider	New Orders	Change Orders
Provider A	30	90
Provider B	60	30

The average time to complete an order for each of the two providers would be computed as follows:

Mean OCI Provider A = [30(2) + 90(1)]/120 = 150/120 = 1.25
Mean OCI Provider B = [60(2) + 30(1)]/90 = 150/90 = 1.67

The seeming discrepancy in the means here is due entirely to the difference in volume of orders under each category. The discrepancy is not due to any difference in time required, because the time required for both providers is exactly the same.

We have attempted to adjust the data to account for differences in volume of orders between BellSouth and the CLECs in aggregate. The distribution of the BST cases is adjusted so that it is more similar to the CLEC distribution, by type of order and location. If, in fact, there is no difference in the distributions (i.e., if they each have the same proportion of new orders, etc.), then the adjusted mean equals the unadjusted. The specific adjustments are detailed in Appendix B. The overall means for the adjusted data are given below:

Figure 3 - Order Completion Interval
Difference in Means, Adjusted Data (August)

Service Provider	Mean (Days)
BellSouth	1.48
CLEC	1.62
Difference	-0.15

The original data show a difference in mean Order Completion Interval of -0.42 days. This magnitude of difference seems, on its face, unacceptable. However, after adjusting

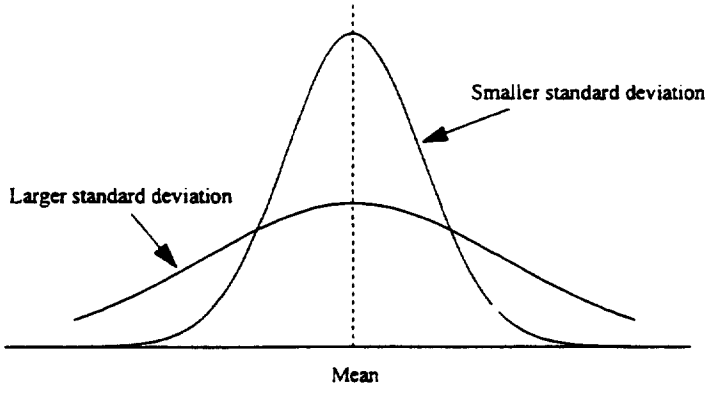
the data, the difference is only -0.15 days (-0.14 if the subtraction is made with the rounded numbers in Figure 3).

Is this likewise unacceptable? A statistical test can be performed to determine whether the difference is statistically significant. But this is not the same as determining whether the difference is of practical importance. A statistically significant difference does not automatically translate into a practically significant difference. Will a customer even notice a difference of 0.15 days in the fulfillment of his or her order? Where is the boundary between unimportant and important? This is a gray area that requires further discussion before judgment can be made. Perhaps a customer survey is needed for these differences that are statistically significant but "too close to call" from a practical standpoint.

Overall CLEC and BellSouth Standard Deviations. – The standard deviation (see Glossary) is another well-known descriptive measure, indicating how the data are spread about the mean. The larger the standard deviation, the more spread out the data. The illustration in the box below points out how data which appear alike when examining only means can actually be different. Standard deviations are most useful for comparison purposes when data are symmetric and normally distributed.

Illustration B – Empirical Example for Interpreting the Standard Deviation

The means of two sets of data can be equal, even while the underlying data used to compute the means are quite different. One set of data may have a larger standard deviation than the other. This graph provides a representation of two different sets of data. Note that the means are equal, but the spread about the means is quite different.



The standard deviations of the underlying data provided in Figure 2 are given in Figure 4, which follows. Notice that the spread about the mean is somewhat larger for BellSouth than for the CLECs. This implies that, as far as the standard deviation is concerned, BellSouth is offering poorer service to its own retail customers than it is to its CLEC resale customers. This difference is clear in the unadjusted data, but becomes even more pronounced after adjustment.

Figure 4 - Order Completion Interval Standard Deviation (August)

Service Provider	Unadjusted	Adjusted
BellSouth	2.78	2.95
CLEC	2.26	2.26
Difference	0.52	0.69

The FCC measure, which pools the standard deviations, is based on an underlying assumption that the standard deviations are equal. The large difference in standard deviations in this case is an indication that the FCC measure may not be suitable for the current situation.

The standard deviation can be computed for any set of data, but it is most meaningful when data are symmetric. For example, the graphical plots in Illustration B show data as normally distributed. When the data are not normally distributed, then the standard deviation still provides a measure of the spread about the mean, but its utility is less clear. Therefore, it is important to check the 'shape' of the data in other ways too, especially when testing for statistical significance.

Overall CLEC and BellSouth Relative Frequency Distributions. – In Figures 5 and 6 below, we check for differences in the distributions in a more graphical way than is afforded by just looking at the value of the standard deviations. For these figures, the horizontal axis is the Order Completion Interval, in days. The vertical axis shows the relative frequency, or percentage of observations at each value. The relative frequency distributions in Figures 5 and 6 show the shape of the unadjusted and adjusted data, respectively, for both BellSouth and the aggregate of all CLEC activity.

Figure 5 - Order Completion Interval
Relative Frequency Distribution of Unadjusted Data (August)

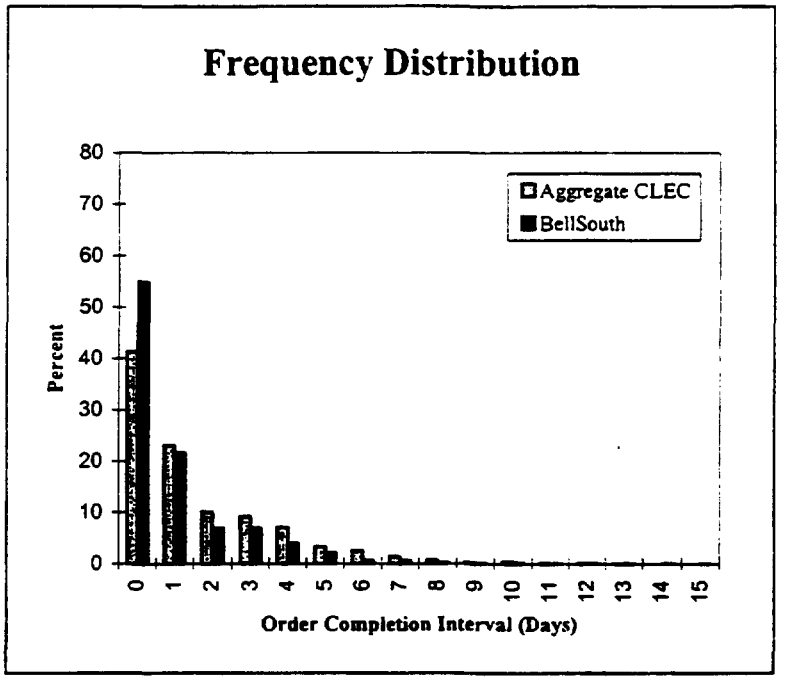
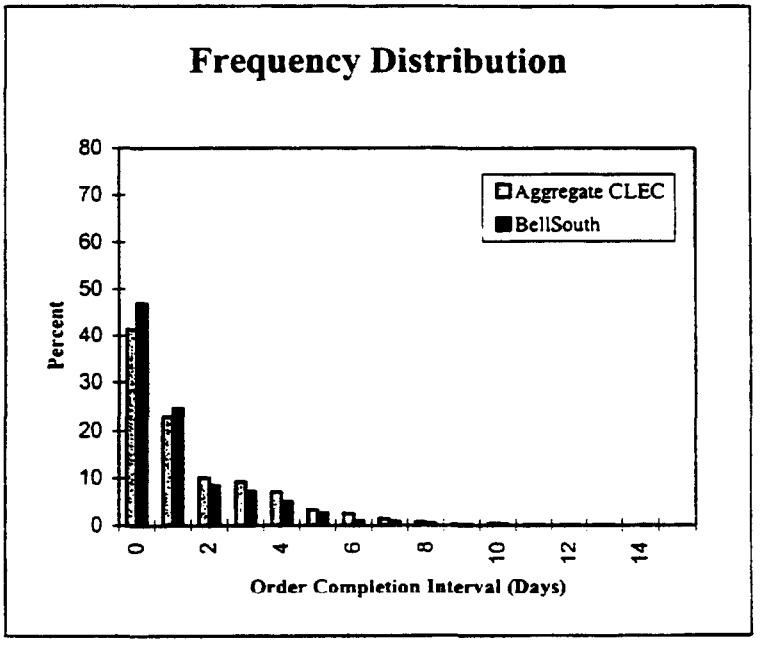


Figure 6 - Order Completion Interval
Relative Frequency Distribution of Adjusted Data (August)



The first thing to note from these graphs is that the data do not look normal or symmetric. More of the observations fall to the left side of the graph. For example, from Figure 5 it can be seen that over 75 percent of the observations lie to the left of the mean for BellSouth unadjusted. Although a smaller percentage lie to the right, the values there can actually be quite large, as high as 15 days in some cases. This type of distribution is referred to as skewed and heavily tailed.

The frequency distributions appear similar between BellSouth and the CLECs, particularly for the adjusted data. Each display a heavily skewed distribution, and the spread is not noticeably different. Given the similar appearance of these distributions, it seems plausible that testing the mean difference between BellSouth and the aggregate CLECs is a reasonable approach to checking for statistical significance between BST and CLECs service.

Overall CLEC and BellSouth Quantile Comparisons. – One way to help the eye interpret the similarity in the relative frequency distributions between BellSouth and CLEC in Figures 5 and 6 is a quantile-by-quantile comparison. The quantile comparison is another descriptive tool (see Illustration C below and Glossary). In Figures 7 and 8, the quantile comparisons of unadjusted and adjusted data are provided.

Illustration C—How to Read a Quantile Plot

A quantile is the value of the distribution at a particular percentile. For example, the quantile associated with the 10^{th} percentile, is the value of the distribution such that 10 percent of the distribution lies at or below that value and 90 percent of the distribution lies above that value. To produce a quantile comparison plot, quantiles from each distribution are computed and plotted against each other. If the distributions are similar, the quantile plot should fall roughly along the 45 degree line of equality. The quantile plot matches the distributions in the tails, not just in the center, and captures the spread of the data. In a sense, it provides a visual summary of all the other measures we have looked at.

The horizontal axis represents the quantile from the CLECs data, while the vertical axis represents the quantile from the BellSouth data. When the two quantiles match, the point on the graph will lie exactly on the 45 degree line of equality. A preponderance of points below the 45 degree line suggest shorter completion times for BellSouth while if most points fall above the line, it suggests shorter completion times for the CLECs.

Figure 7 - Order Completion Interval
Quantile Comparison of Unadjusted Data (August)

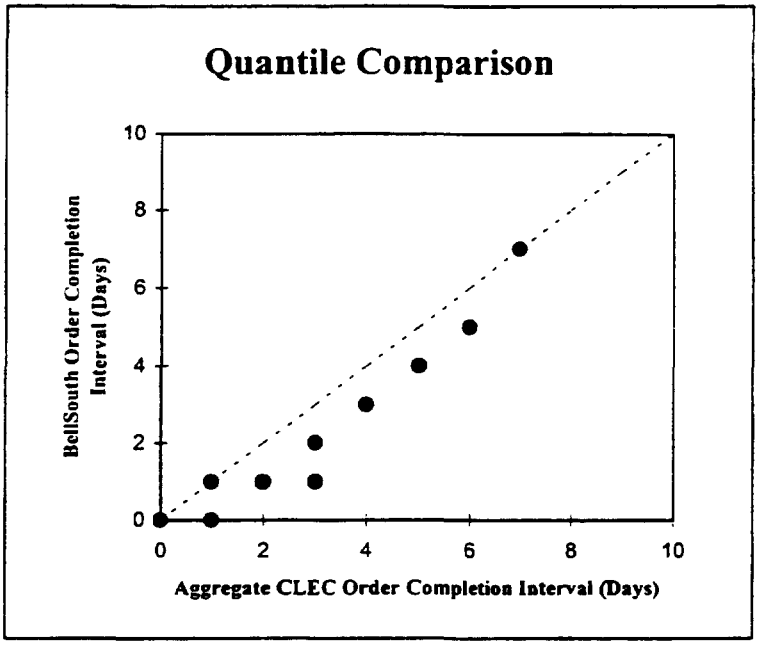
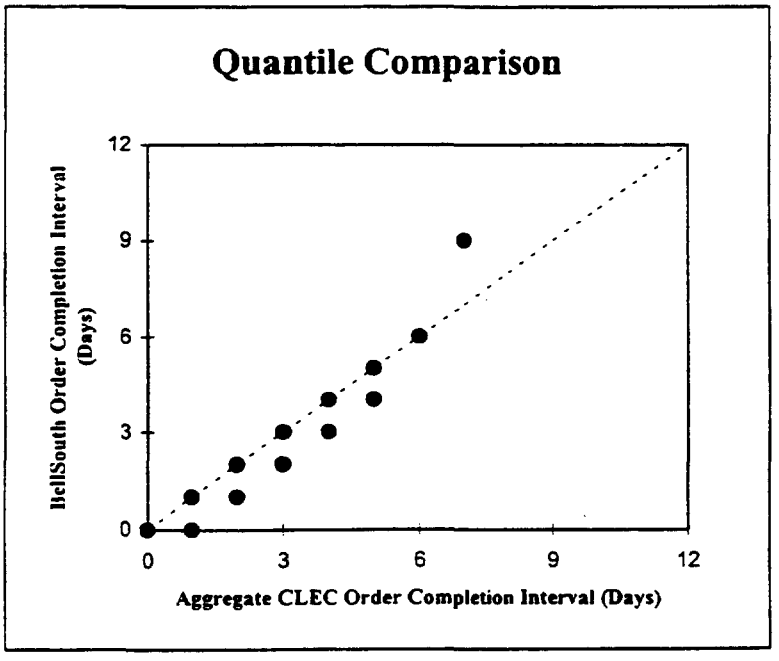


Figure 8 - Order Completion Interval
Quantile Comparison of Adjusted Data (August)



The quantile comparisons indicate that for the unadjusted data, BellSouth and CLEC do not seem to follow the same distributions. For the adjusted data, the distributions are more similar, albeit still not identical. For example, the BellSouth adjusted data appears to have a slightly heavier tail than for CLECs (as indicated by the last point in Figure 8 which lies well above the 45 degree line of equal treatment).

Previously Proposed Test Statistics. – The final tools employed here are the set of three statistical tests that the Commission wanted us to compare. It is important to understand the differences between the tests. Each is a test of the difference in means between BellSouth and the CLECs. Each builds upon a standard formula for comparison of two means: the numerator is the difference in the means, the denominator is a standard error estimate. Assumptions about the test statistics vary and are covered in depth in Appendix B.

- LCUG Test Statistic: The LCUG Modified Z assumes that the best measure of the standard error of the difference is a function of the BellSouth data only. Since the BellSouth data have many more observations than the CLEC data, the idea is that, if the CLEC and BellSouth data come from a common distribution, most of the information about variation in the data is coming from BellSouth anyway. Therefore, why not simply reflect this directly?

- FCC Test Statistic: The FCC Pooled Z, unlike the LCUG measure, assumes that the BellSouth and CLEC variances can be pooled. That is, the standard deviation used in the denominator of the test is a weighted combination of the standard deviations from each set of data.

CLEC and BellSouth Plots to Check Assumptions. --Both the LCUG Modified Z and the FCC Pooled Z statistics require the assumption of independence (see Glossary) between the observations and the assumption that each observation comes from the same distribution. There is evidence that the first assumption is not satisfied. In fact, there appears to be a great deal of dependence in the data. Figure 9 below shows a scatterplot of the means for BellSouth and CLEC at the Wire Center level for the unadjusted data. Figure 10 shows the same for the adjusted data. The line superimposed on the scatterplot was found using least trimmed squares regression (see glossary). This method differs from ordinary regression in that it guards against extreme values influencing the outcome.[1]

[1] A straight line is not necessarily the appropriate function to use to fit this data. It does, however, illustrate the dependence structure contained in the data.

Figure 9 - Order Completion Interval
Scatterplot of Means by Wire Center, Unadjusted Data (August)

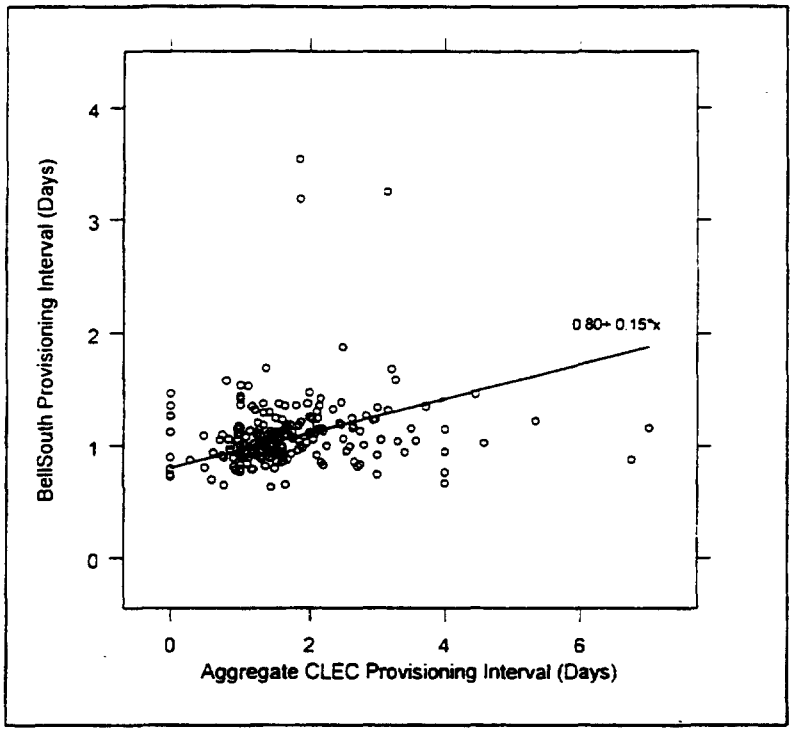
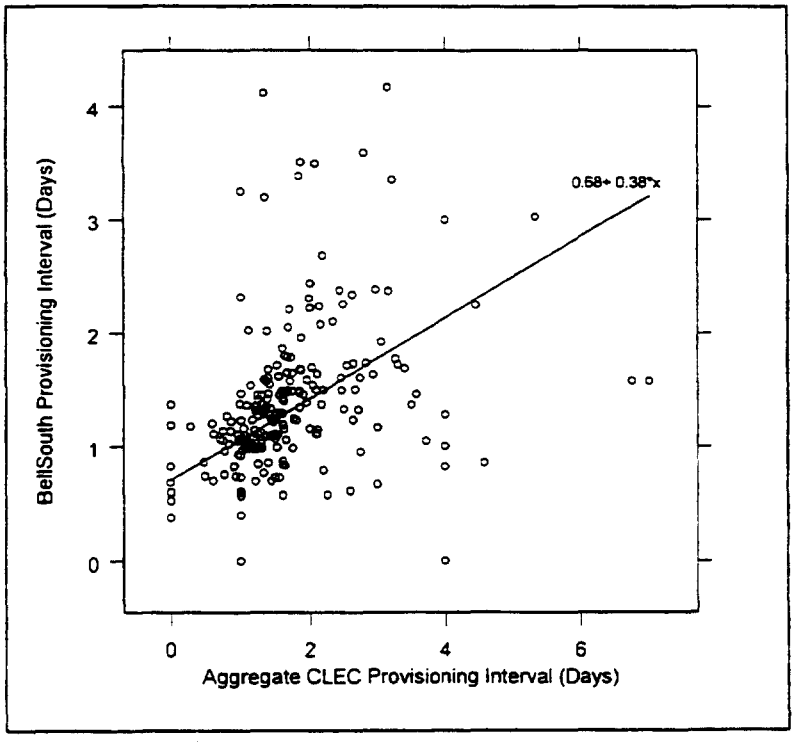


Figure 10 - Order Completion Interval
Scatterplot of Means by Wire Center, Adjusted Data (August)



If there were little or no dependence in the data, then the line through the data in Figures 9 and 10 would be flat. This is not at all the case, however, especially after adjusting the BellSouth data so that its relative volumes parallel those of all the CLECs combined.

BellSouth Proposed Test Statistic. – The third test statistic examined is an alternative to the two discussed above. This test statistic is designed to compare "like-to-like," to capture differences in the standard deviations between BellSouth and the CLECs, and to be robust against failures in the assumption of independence. Our approach is based on a well-known technique for variance estimation called the random group or "replicate method." In the current application, the method breaks the data up into separate sets, or replicates, of wire centers so that each set of wire centers can then be treated as approximately independent and identically distributed.

For our proposed test, the data were divided into 30 replicates and the differences and standard deviations were calculated from the replicates. The standard t-test was then applied to any differences found. The advantage of this approach is that it better reflects the dependence within the data. It is conceivable that there are further dependencies that have not been captured by the structure of the replicates. If so, this test statistic may be slightly biased against BellSouth. The method is described in full technical detail in Appendix B.

Results of Statistical Tests. – Figures 11 and 12 below present the results from each of the three statistical testing methods discussed: the LCUG, FCC, and BellSouth tests. There are two columns in each figure; the first column presents the value of the test statistic itself, and the second column presents the associated P-values.

Figure 11 - Order Completion Interval
Test Statistics, Unadjusted Data (August)

Testing Method	Test Statistic	P-value (percent)
LCUG	-18.70	0.0000
FCC	-18.83	0.0000
BellSouth	-9.37	0.0000

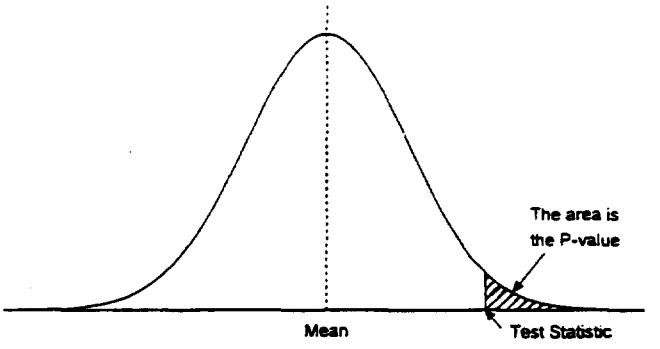
If we consider for the moment the results for the unadjusted data in Figure 11, we see large values for the test statistics. The minus signs before the values indicate that this difference favors BellSouth. The P-value indicates how extreme the test statistic is. The description box (Illustration D) explains a bit more about what the P-value means. In each of the three cases here, the P-value indicates that there is a statistically significant difference in means.

Illustration D – Graphical Interpretation of the P-Value

If we have a curve depicting the distribution of our test statistic, the P-value is the area under the curve to the more extreme side of the statistic.

If the statistic is positive, the P-value is the area to the right. If the statistic is negative, the P-value is the area to the left.

If the P-value is greater than 2.5%, or 0.025, then in this report we treat the test results as not yielding what will be considered statistically significant.



Another point to notice is that while the LCUG and FCC statistics are very close to each other, the statistic from the proposed method is significantly smaller, in terms of absolute value. This simply demonstrates that there are different ways of thinking about the data and what it says. Once again, the BellSouth method is designed to account for inherent dependencies within the data and for differing standard deviations between BellSouth data and the CLECs.

Figure 12 - Order Completion Interval
Test Statistics, Adjusted Data (August)

Testing Method	Test Statistic	P-value (percent)
LCUG	-6.08	0.0000
FCC	-6.13	0.0000
BellSouth	-2.57	0.7774

As discussed already, the goal of adjusting the data is to compare "likes-to-likes." So far we have only looked at results for the unadjusted data. If we now look at the results for the adjusted data in Figure 12, and compare them to Figure 11 (unadjusted), we see quite a large difference in the test statistics. Even so, the basic story has not changed: the statistics are all negative, showing a difference in favor of BellSouth, and the P-values all show that these differences are statistically significant. Again, the LCUG and FCC

statistics are close to each other, while the statistic from the BellSouth method is less than half of the others. The difference between the results for the adjusted "like-to-like" data and those for the unadjusted data is in the magnitude of the test statistics. The adjusted data yield statistics that are about one third of what they are for unadjusted data, indicating the importance of performing such adjustments so as to assure the analysis is appropriate.

III. Maintenance - Maintenance Average Duration

Maintenance Average Duration is another measure that has intuitive appeal for those seeking to investigate differences between BellSouth and the CLECs. Maintenance Average Duration measures the amount of time, in hours, that it takes for a trouble[1] to be resolved. It seems reasonable that if both BellSouth and CLECs are receiving the same quality treatment, on average their respective troubles should be resolved in about the same amount of time.

As with Order Completion Interval, the Provisioning measure analyzed in Section II, Maintenance Average Duration is reported at a number of levels in an effort to group together similar types of troubles for comparison. Troubles are separated as Designed/Non-Designed, Dispatch/Non-Dispatch, and, for Non-Designed troubles, Residence/Business (Designed troubles are not distinguished as Residence or Business). Figure 13 is an example from the August report for Maintenance Average Duration already provided to the commission.

Figure 13 - Extract of August Maintenance Average Duration Report

RESALE SERVICES - RESELLER: AGG - CLEC Aggregate

Report Period: 08/01/1998 to 08/31/1998

SQM: Maintenance Average Duration
Non-detailed Report

	Residence			Business		
	Dispatched	Non-Disp.	Total	Dispatched	Non-Disp.	Total
ALABAMA	36.71	9.40	29.77	14.61	9.79	12.89
FLORIDA	26.53	12.08	20.97	18.84	12.55	16.04
GEORGIA	28.51	14.37	24.00	14.35	7.60	11.79
KENTUCKY	28.58	14.63	25.21	21.58	10.69	17.49
LOUISIANA	36.77	11.80	30.90	21.29	9.47	16.88
MISSISSIPPI	37.11	9.10	27.71	13.97	1.74	10.91

As with Order Completion Interval, there are different ways one can look at and analyze the average duration results to determine what the data are telling us. Once again, it is desirable to have an overall view of the data so that a combined comparison can be made. We again present a variety of tables and graphical displays in order to better visualize the distribution and similarity or dissimilarity of the data, and to enable more appropriate assumptions to be made. These assumptions have an effect on the choice of statistical test. Following these tables and graphs, we present the results of the three different

[1] It should be noted that whereas an individual data record for Order Completion Interval is referred to as an 'order', for Maintenance Average Duration the more appropriate term is 'trouble.' This is the term that will be used throughout this Section.

statistical tests, the same tests that were performed for Order Completion Interval. What follows will closely parallel the discussion of OCI.

At the time of this report, we have only received complete data for the Non-Design portion of this measure. Therefore, all graphs and tables that follow are derived solely from the Non-Design data.

As with OCI, the Maintenance Average Duration data cannot necessarily be analyzed accurately in their raw form. Largely disparate volumes of troubles received for BellSouth and for the CLECs is evidence to suggest that the data must be adjusted before valid analyses can be made. Accordingly, the following tables and graphs present both unadjusted and adjusted data. The adjustments performed on this data are described in Appendix B.

Overall CLEC and BellSouth Means. – Once again, the unadjusted mean is computed from the original BellSouth and CLEC data. It is consistent with the reports provided to the Commission earlier (see Appendices E and F). The adjusted mean is computed from data that has been standardized to reflect relative differences in volume between CLECs resale maintenance requests and their corresponding BellSouth retail maintenance requests. Figure 14 summarizes these means and their differences.

Figure 14 - Maintenance Average Duration, Non-Design Differences in Means (August)

Service Provider	Mean Hours (Unadjusted)	Mean Hours (Adjusted)
BellSouth	23.45	26.51
CLEC	27.89	27.89
Difference	-4.44	-1.38

As can be observed from Figure 14 for the unadjusted data, the CLECs clearly have a higher mean than does BellSouth. After the adjustment to the data, the discrepancy is much smaller, little more than a third of what it at first appeared.

Once again the question is how big are these differences from a practical standpoint? Are they even statistically significant? If they are, are they practically significant? It seems that a difference of nearly four and a half hours would be an important difference indeed. But what about 1.4 hours? In the context of a task that requires on average about 27 hours to complete, will an hour and 25 minutes make a difference? These questions on the practical significance of the difference can likely only be answered by the customers themselves.

Overall CLEC and BellSouth Standard Deviations. – The standard deviations help give us an idea of how the data are spread out about the mean. Again, we must keep in mind that the standard deviation is useful for comparison purposes mainly when data are

symmetric and normally distributed. Figure 15 presents the standard deviations for the unadjusted and adjusted Maintenance Average Duration data so that statistical tests of the means can be interpreted properly.

Figure 15 - Maintenance Average Duration, Non-Design Standard Deviations (August)

Service Provider	Unadjusted	Adjusted
BellSouth	25.18	27.05
CLEC	27.48	27.48
Difference	-2.30	-0.43

We see from Figure 15 that the standard deviation is slightly larger for the CLECs than it is for BellSouth. The adjustment increased the standard deviation for BellSouth, and thus the discrepancy between the BellSouth and CLEC standard deviation is almost eliminated.

Overall CLEC and BellSouth Relative Frequency Distributions. – For Maintenance Average Duration, the horizontal axis of the frequency distribution represents the time taken to resolve an individual trouble, in hours. The vertical axis shows the relative frequency – the percentage of all troubles – in that range. The relative frequency distributions in Figures 16 and 17 show the shape of the unadjusted and adjusted data, respectively, for both BellSouth and CLEC. Once again, the purpose of looking at these frequency distributions is to check for differences in the distributions of the datasets so that statistical tests of the means can be interpreted properly.

Figure 16 - Maintenance Average Duration, Non-Designed Relative Frequency Distribution of Unadjusted Data (August)

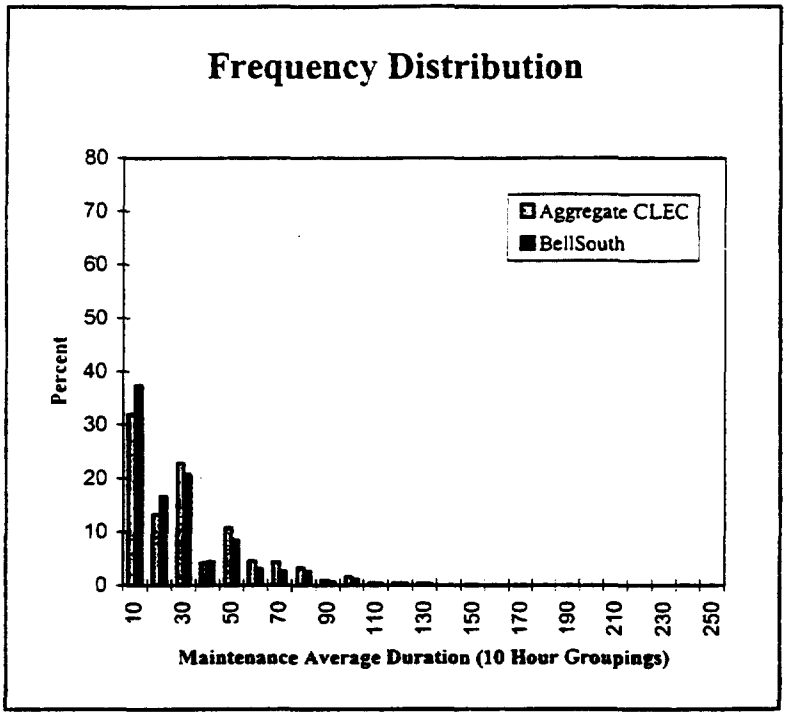
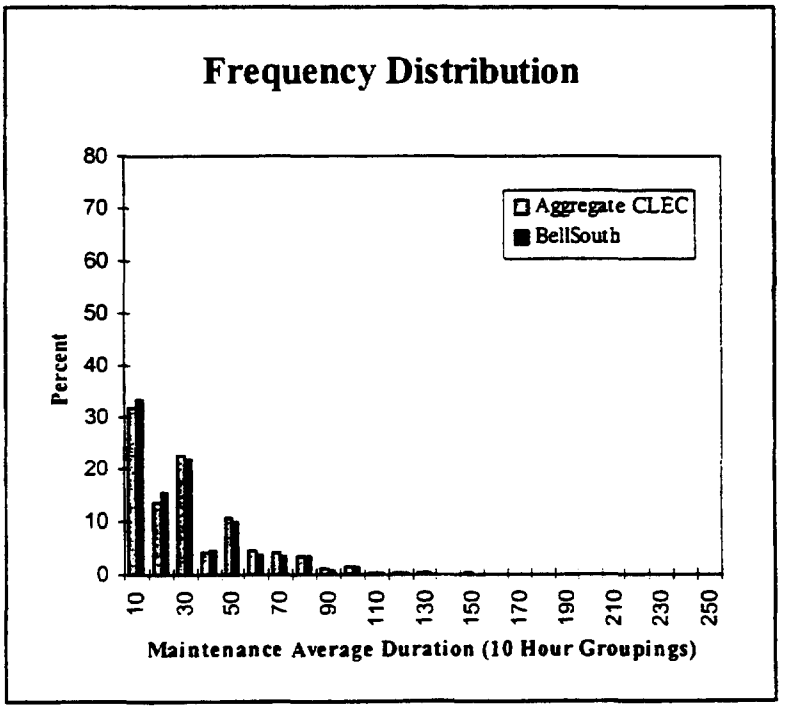


Figure 17 - Maintenance Average Duration, Non-Designed Relative Frequency Distribution of Adjusted Data (August)



Although the distributions for the unadjusted data were fairly similar to begin with, the adjustment to the data equalized the CLEC and BellSouth distributions to an even greater degree. The graphs both have a spike in the 10-hour grouping, another in the 30-hour grouping, and another, smaller spike in the 50-hour grouping, and then trail away after that. Because the distributions are so similar, especially for the adjusted data, a statistical test of the difference of means may be a reasonable approach to investigating whether BST and CLECs service are substantially similar on this measure.

Overall CLEC and BellSouth Quantile Comparisons. – Figures 18 and 19 present the quantile comparisons for the unadjusted and adjusted Maintenance Average Duration data, respectively. Just as with OCI, the horizontal axis represents the quantile from the CLEC data while the vertical axis represents the same quantile from the BellSouth data. If the value at the 10^{th} percentile of the BellSouth data is the same as the value at the 10^{th} percentile of the CLEC data, then the point will fall on the 45-degree line.

Figure 18 - Maintenance Average Duration, Non-Designed Quantile Comparison of Unadjusted Data (August)

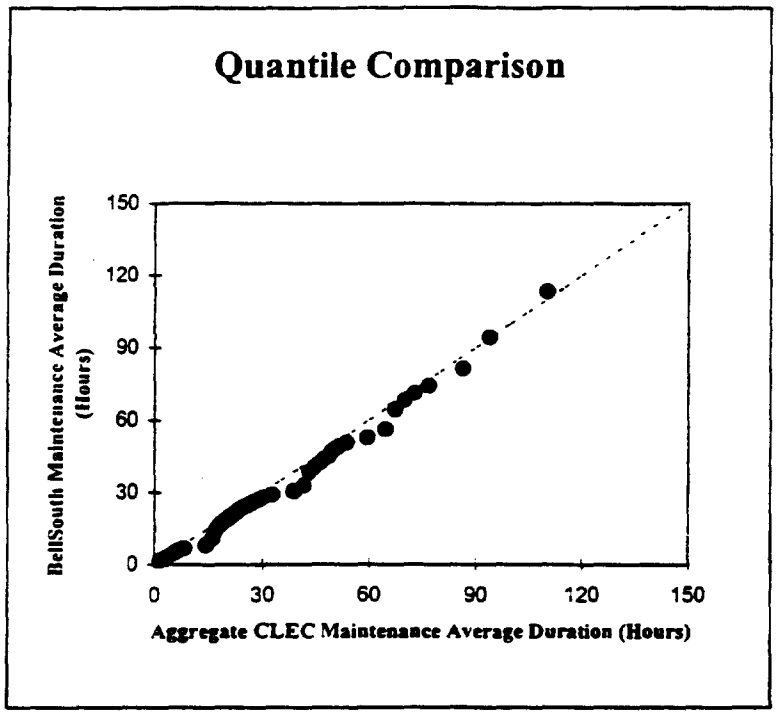
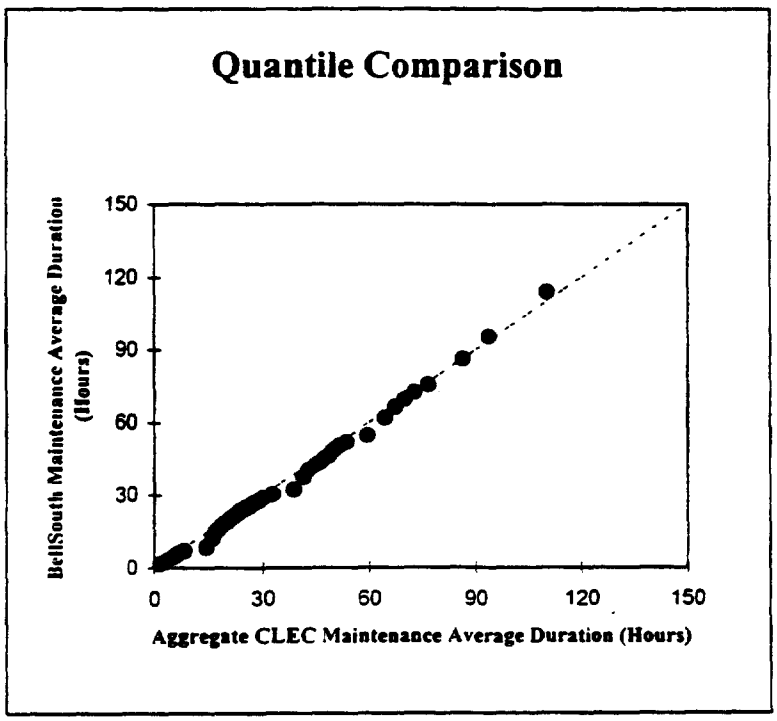


Figure 19 - Maintenance Average Duration, Non-Designed Quantile Comparison of Adjusted Data (August)



These quantile charts reinforce what we learned from the frequency distributions: That the distributions of the BellSouth and CLEC data sets are really very similar. In the unadjusted graph, Figure 18, the majority of the points lie only slightly under the 45-degree line of equality. In Figure 19 we see that the adjusted data has caused the points to fall a bit more often on the 45 degree line itself.

CLEC and BellSouth Scatterplots to check assumptions. – Both the LCUG Modified Z and FCC Pooled Z statistics require the assumption of independence between the observations. There is evidence that this assumption is not satisfied. In fact, there appears to be a great deal of dependence within the data. The figure below, Figure 20, shows a scatterplot of the unadjusted wire center means for BellSouth and CLEC. Figure 21 shows the same for the adjusted means. Both figures are for August.

Figure 20 - Maintenance Average Duration, Non-Designed
Scatterplot of Means, Unadjusted Data by Wire Center (August)

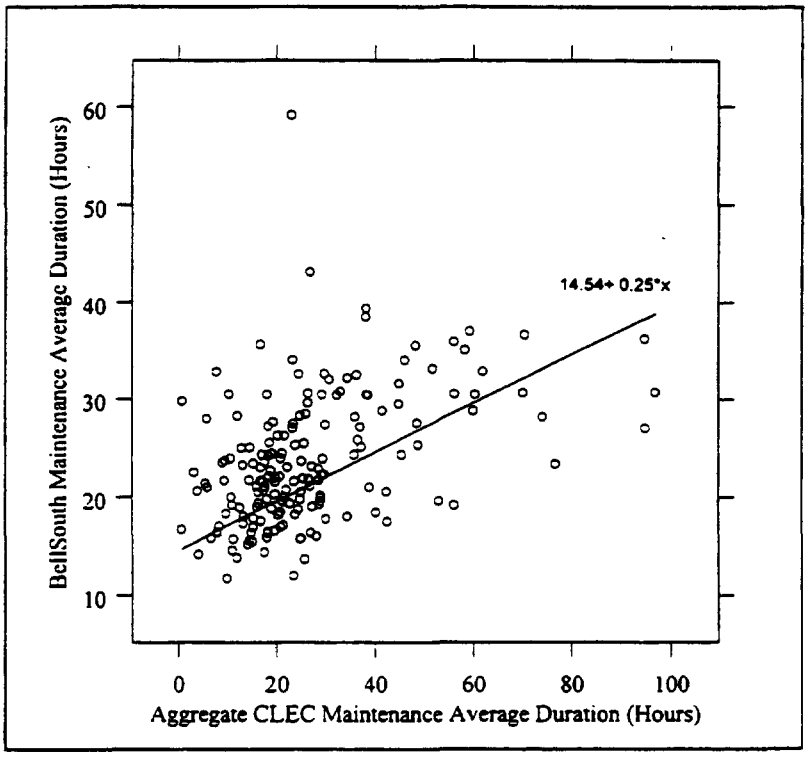
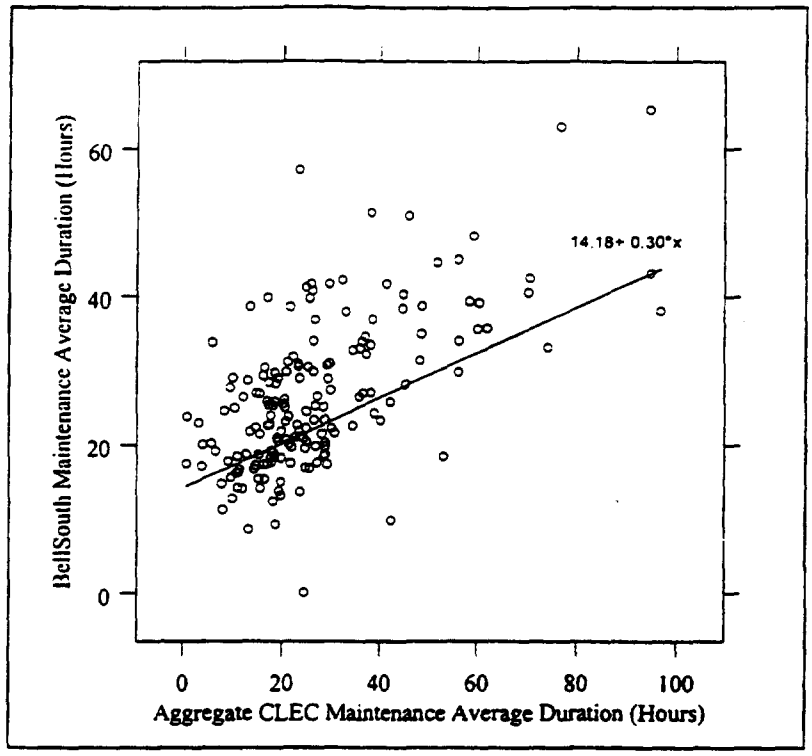


Figure 21 - Maintenance Average Duration, Non-Designed Scatterplot of Means, Adjusted Data [by Wire Center] (August)



If all of the observations were completely independent, we would expect the points on the scatterplots to be randomly disbursed throughout the graph area. As a result, a least trimmed squares (see Glossary) is drawn through the points would have a slope very near to zero, which would be a horizontal line. Both graphs, however, do not have a straight line with a slope near zero. Instead, they both have very definite positive slopes, indicating that a wire center which has a high average duration for the CLECs will tend to have a high average duration for BellSouth.

Results of Statistical Tests. – We have now again reached the point where we present the statistical tests themselves. We repeat that the assumptions underlying each of the different test statistics vary (see Appendix B). We should look for the technique which makes assumptions that are most consistent with the results of the tabular and graphical analyses just presented.

The three tests have been briefly described in the Order Completion Interval Section, and again they are described in full detail in Appendix B. Below them are the three sets of test results for Maintenance Average Duration.

Figure 22 - Maintenance Average Duration, Non-Designed
Test Statistics (August)

	Unadjusted		Adjusted	
Testing Method	**Test Statistic**	**P-value (percent)**	**Test Statistic**	**P-value (percent)**
LCUG	-6.62	0.0000	-1.91	2.7770
FCC	-6.61	0.0000	-1.91	2.7809
BellSouth	-3.28	0.1356	-1.93	3.1656

If we first look at the results for the unadjusted data, we see that the test statistics for the LCUG method and for the FCC method are nearly equal. The BellSouth test statistic is about half of these. This is quite similar to the situation we found for the Order Completion Interval. The P-value for each test indicates that there is a strongly statistically significant difference in the unadjusted means. Moreover, because the statistics' signs are negative, that difference favors BellSouth.

The same statistics for the adjusted data are quite a bit lower, in terms of absolute value. In fact, the new LCUG and FCC statistics are less than a third of their unadjusted counterparts. This time, all three are close to each other. And the differences in the means, it could be argued, are borderline significant at worst.

The main story from these test statistics is that, initially, there was a strong indication of a difference in the overall unadjusted means for Maintenance Average Duration between the CLECs and BellSouth. That indication, though still present for August, is not at all strong after the data is adjusted to account for different relative volumes of observations. When adjusted, the three different statistical tests yield quite similar results. The need to choose a statistical test which employs assumptions that are in accord with the true characteristics of the data, as demonstrated by the graphs and tables presented above, cannot be overstated.

IV. Operating Support Services (OSS) Response Interval

Different in nature from the Order Completion Interval and Maintenance Average Duration, the manner in which the OSS Response Interval data is stored limits the level of analysis that can performed on it. While the two aforementioned performance measures allow for the breakdown of data by service categories (e.g., Dispatched vs. Non-Dispatched), no such breakdown exists for OSS Response Interval.

Two negotiation systems act as interfaces, which allow BellSouth and the CLECs to perform preordering functions electronically without assistance or intervention from BellSouth personnel. The Regional Negotiation System (RNS) handles BellSouth requests, while the Local Exchange Negotiation System (LENS) is employed by the CLECs. Each system allows users to obtain information from a number of BellSouth operating systems and corporate databases. The amount of call time in milliseconds and the number of calls make up the bulk of the applicable information that is stored daily for the calculation of an average response interval.

Without the knowledge of the length of each individual call, we are unable to calculate a variance for the average response intervals. Thus, none of the statistical tests employed earlier can be used here. As a result, an alternative method must be derived to statistically compare the average response intervals.

Concentrating on a three month period from July to September 1998, daily OSS Response Interval data were available for 13 systems. Six systems are available only in RNS, and thus used exclusively by BellSouth, while three systems are available only in LENS and used only by the CLECs. Four systems (ATLAS, DSAP, RSAG (By ADDR) and RSAG (By TN)) are available in both RNS and LENS. Those systems are as follows:

- **Application for Telephone Number Load Administration and Selection (ATLAS)** - The BellSouth operating system used to administer the pool of available telephone numbers and to reserve selected numbers from the pool for use on pending service requests/service orders.

- **DOE Support Application (DSAP)** - The BellSouth operating system which assists a Service Representative or similar carrier agent in negotiating service provisioning commitments for non-designed services and UNEs.

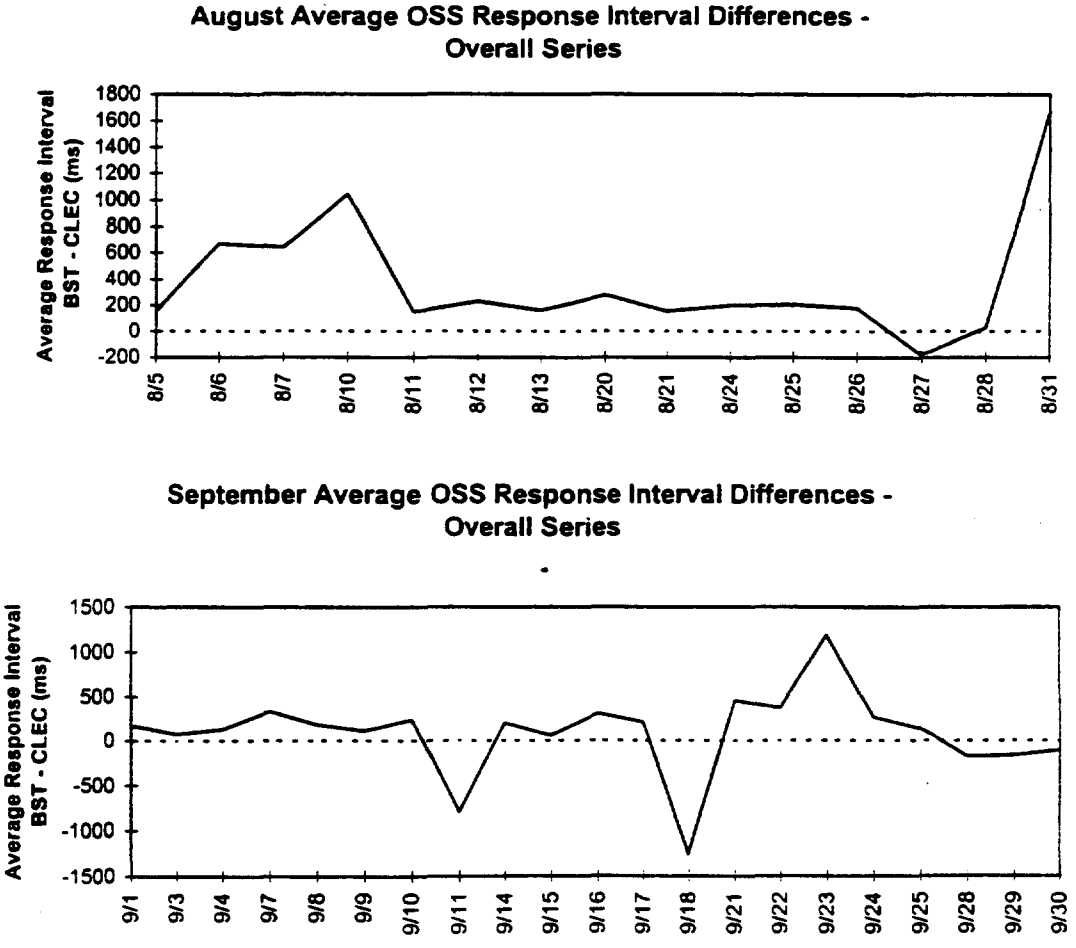
- **Regional Street Address Guide (RSAG(By ADDR))** - The BellSouth database which contains street addresses validated to be accurate with state and local governments. RSAG (By ADDR) refers to requests based on address

- **Regional Street Address Guide (RSAG(By TN))** - The BellSouth database which contains street addresses validated to be accurate with state and local governments. RSAG (By TN) refers to requests based on telephone number.

To compare the average response interval for BellSouth to the CLECs, we limited our analysis to the four systems for which there were "like-to-like" data. For each day for which data existed, we determined a daily average response interval by taking the total amount of call time and dividing it by the number of calls. The CLEC daily average response intervals were subtracted from the corresponding BellSouth intervals, yielding a series of daily average response interval differences. An overall series was also calculated as an average difference over the four sets of daily average response interval data.

Concerned with the possibility of a time dependence within the data, we employed time series analysis methodology. Figure 23 illustrates the average response interval differences for the overall series for both August and September.

Figure 23 - Time Series of Average Response Interval Differences (Overall)



Looking at the individual values for both August and September for the average response interval differences, we see that almost all of the differences are positive, thus denoting a longer daily average response interval for BellSouth. Of the fifteen days in August in which data were collected, a negative difference was recorded for only one day. For

September, of the twenty-one days in which data were collected, a negative difference was recorded for only two days.

Common statistical procedures and techniques often rely on the assumption that observations are independent and identically distributed. The intrinsic nature of time-ordered data is that observations may be dependent or correlated, and the order of the observations is, therefore, important. This potential dependence is similar in nature to geographical data, in which observations may be dependent upon location, as discussed elsewhere (see especially Appendix B) for Order Completion Interval and Maintenance Average Duration.

We focused our attention on the selection of a statistic to test for disparity between the BellSouth and the CLECs average response intervals. Time series analysis methodology allowed us to investigate the issue of dependence within the data, and use this knowledge to circumvented the difficulties in finding an overall variance in the data. (See Appendix G for the details.)

Figure 24 illustrates the test results for the overall data. Tests were performed for each system and the overall data by month and P-values were calculated based on the degrees of freedom (df) of each test, the results of which are presented in Appendix G.

Figure 24 - OSS Reponse Interval Test Statistics, Overall Data

Overall

Month	Test Statistic	P-value (percent)
July	0.5396	29.7446
August	3.7770	0.0592
September	1.2031	12.1163

The test checks to see if the mean of the daily differences is equal to zero, that is to say that the average response intervals are equal for both BellSouth and the CLECs. A positive test value suggests that BellSouth has a larger average response interval, while a negative test value suggests the converse. Based on the magnitude of the test statistic value and the number of observations employed in the calculation, a P-value can be derived. The results in Figure 24 illustrate that for August, BellSouth is favoring the CLECs.

V. Geographic Diaggregation

In this section, we provide a geographical analysis for the Order Completion Interval. After understanding the general approach and results given here, it may also be useful to go back and compare these results to those obtained for the overall data in Section II.

First it makes sense, however, to discuss why we chose the four local access transport areas (LATAs) as the geographic units to employ. In particular, what about using metropolitan statistical areas (MSAs)? Our reasoning (see Figure 25) was as follows:

- LATAs are a meaningful geographic business unit for BellSouth. MSAs are statistical entities, subject to redefinition by the Office of Management and Budget (OMB) of the Federal government. In fact, there will be a major revision of MSAs in connection with the 2000 Census. Hence MSA units may not have a stable definition over time

- MSA's in Louisiana vary considerably in size as measured by the number of wire centers servicing them. This has the effect for several of making the sample sizes available "small" – too small to safely employ the types of statistical test discussed in this report.

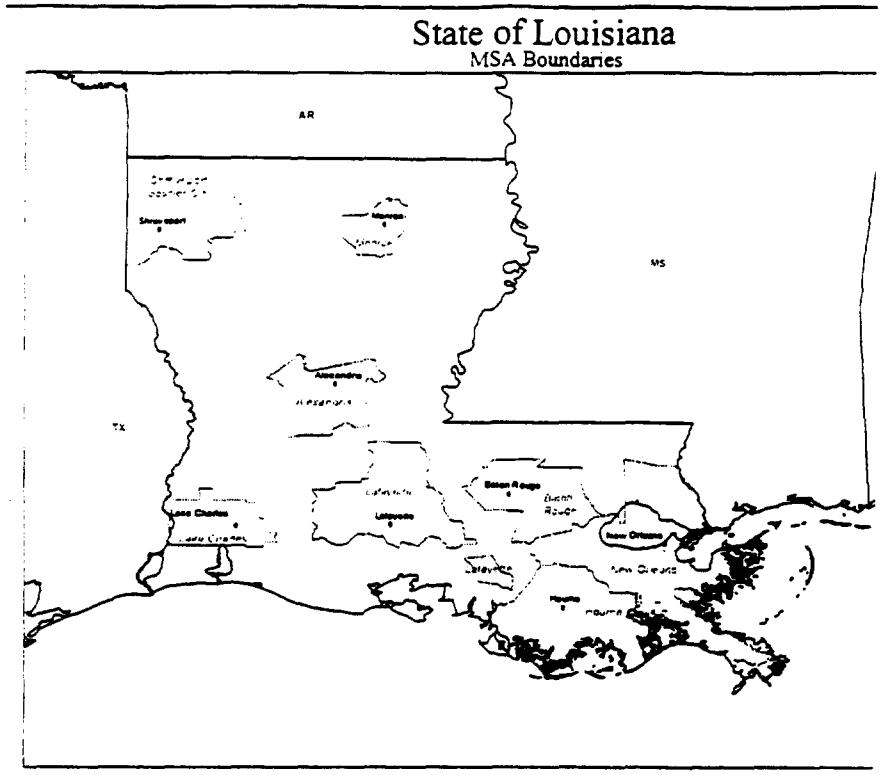
As will be recalled, Section II of this report described the analysis that was performed on the Provisioning measure, Order Completion Interval. It presented a number of tables and graphs describing the August data for that measure, and then presented the results of three formal statistical tests, one advocated by the LCUG, one by the FCC, and one by BellSouth, as required by the Louisiana Order. This section will use the same general format in order to describe the LATA-by-LATA analysis which was performed on the Order Completion Interval measure. The analysis was also performed for the Maintenance Average Duration measure; because the steps are exactly parallel to those presented here for Order Completion Interval, there is not a separate section detailing the latter measure. The charts and tables for the Maintenance Average Duration analysis can be found in Appendices H and I.

Put briefly, examining the data by LATA allows a bit more refinement and isolation of BST and aggregate CLECs differences. For example, we can tell if one region is disproportionately affecting the overall numbers for the state. We can see if the change in averages from one region to the next for the CLECs mirrors that of BellSouth. This is simply a way of reaching further into the data to gain a bit more understanding of what they can really tell us.

Once again, the graphs and tables here will be exclusive to the overall data for August. The entire set of August graphics will be found in Appendix H. The set of September graphics for the LATA analysis are in Appendix I.

Overall CLEC and BellSouth Means. – Figure 26 shows the August breakdowns of the average order completion interval times for BellSouth versus the aggregate CLECs by the four LATA. The means are presented for both the unadjusted and adjusted data.

Figure 25 - Louisiana MSA and LATA Boundaries



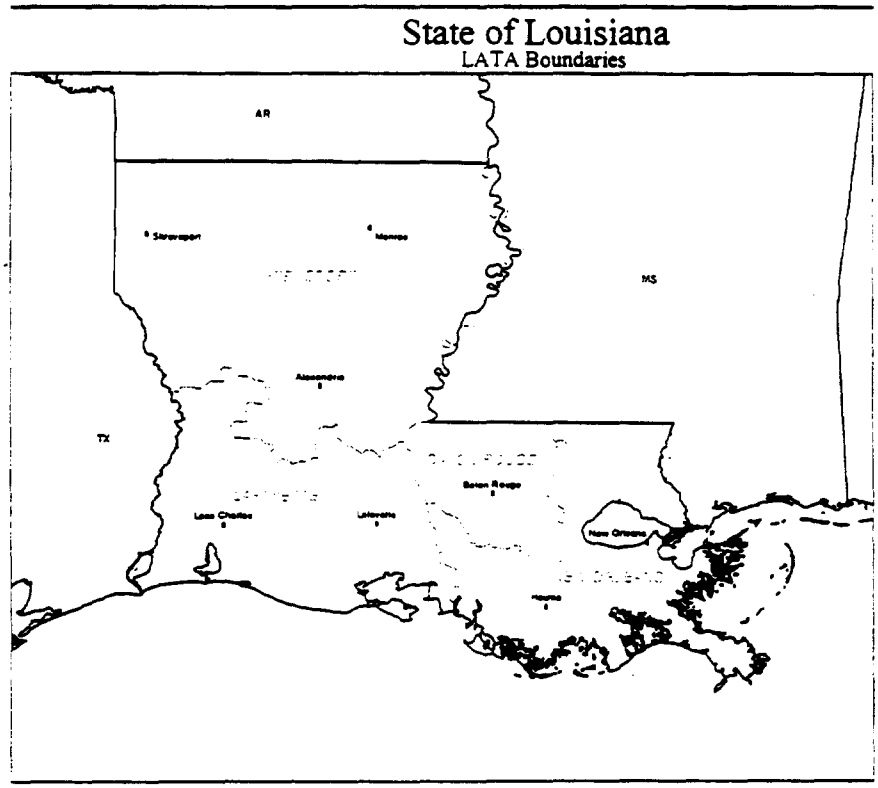


Figure 26 - Order Completion Interval
Differences in Means, by LATA (August)

LATA	Service Provider	Unadjusted (Days)	Adjusted (Days)
Shreveport	**BST**	1.16	1.41
	CLEC	1.82	1.82
	Difference	-0.67	-0.42
Lafayette	**BST**	1.11	1.21
	CLEC	1.38	1.38
	Difference	-0.27	-0.17
New Orleans	**BST**	1.21	1.70
	CLEC	1.57	1.57
	Difference	-0.37	0.12
Baton Rouge	**BST**	1.35	1.44
	CLEC	1.58	1.58
	Difference	-0.24	-0.14

If we look at the means for the unadjusted data first, we see that in each region there is an apparent difference in favor of BellSouth. The average number of days to complete an order for BellSouth is slightly lower in each region than it is to complete an order for the CLECs. This difference in each case is less than a day. The largest difference is seen in the Shreveport LATA and stands at -0.67 days.

When we now look at the adjusted data means, we see that the differences have in each LATA improved to some degree. The largest difference, in terms of absolute value, is still in Shreveport but is now -0.42 days. Interestingly, in the New Orleans LATA, the means actually favor the CLECs after the adjustment. BellSouth is requiring slightly more time, on average, than the CLECs are for that region.

Another point that is perhaps important to notice is that, while for the unadjusted data three of the LATA have differences less than four-tenths of a day, for the adjusted data those same LATA have differences that are less than two-tenths of a day.

Overall CLEC and BellSouth Standard Deviations. – Figure 27 presents the August breakdowns of the standard deviations of the order completion interval times for BellSouth and the CLECs, by LATA. Standard deviations for unadjusted and adjusted data are again presented side-by-side for comparison.

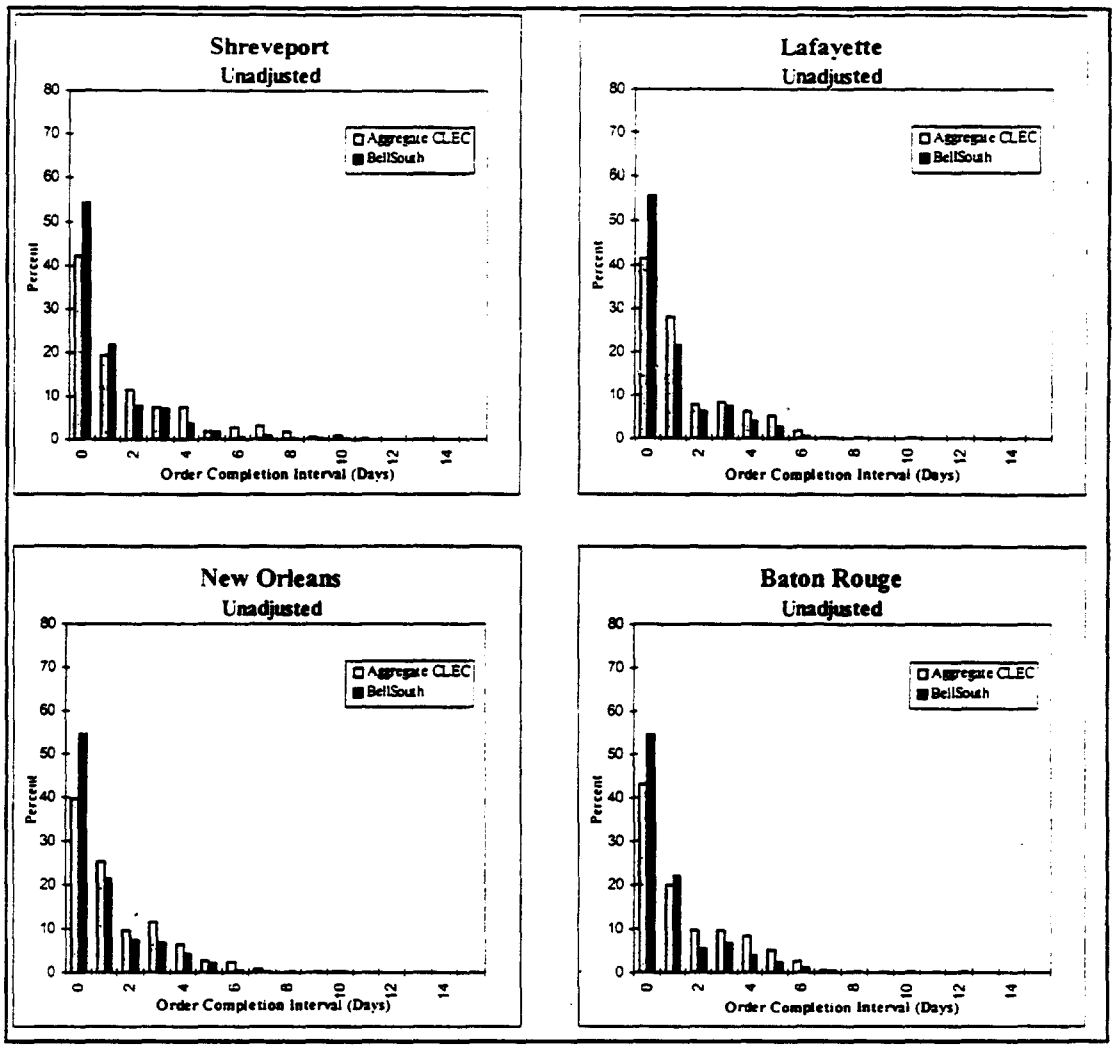
Figure 27 - Order Completion Interval Standard Deviations, by LATA (August)

LATA	Service Provider	Unadjusted (Days)	Adjusted (Days)
Shreveport	**BST**	2.34	2.54
	CLEC	2.54	2.54
Lafayette	**BST**	2.31	2.24
	CLEC	1.71	1.71
New Orleans	**BST**	2.90	3.53
	CLEC	2.25	2.25
Baton Rouge	**BST**	3.41	3.00
	CLEC	2.19	2.19

For the unadjusted data, we see that the standard deviation actually favors the CLECs in each LATA except for Shreveport. In each of the other three, BellSouth has a higher standard deviation, indicating that there is more spread in the BellSouth data. After adjusting the data, the BellSouth standard deviations improve in two of the LATA, and grow worse in the other two. After adjustment, BellSouth and the CLECs have exactly the same standard deviation in Shreveport, which suggests that the pooled variance test of the FCC may be acceptable if restricted to this LATA. The other three LATA show different standard deviations, each favoring the CLECs.

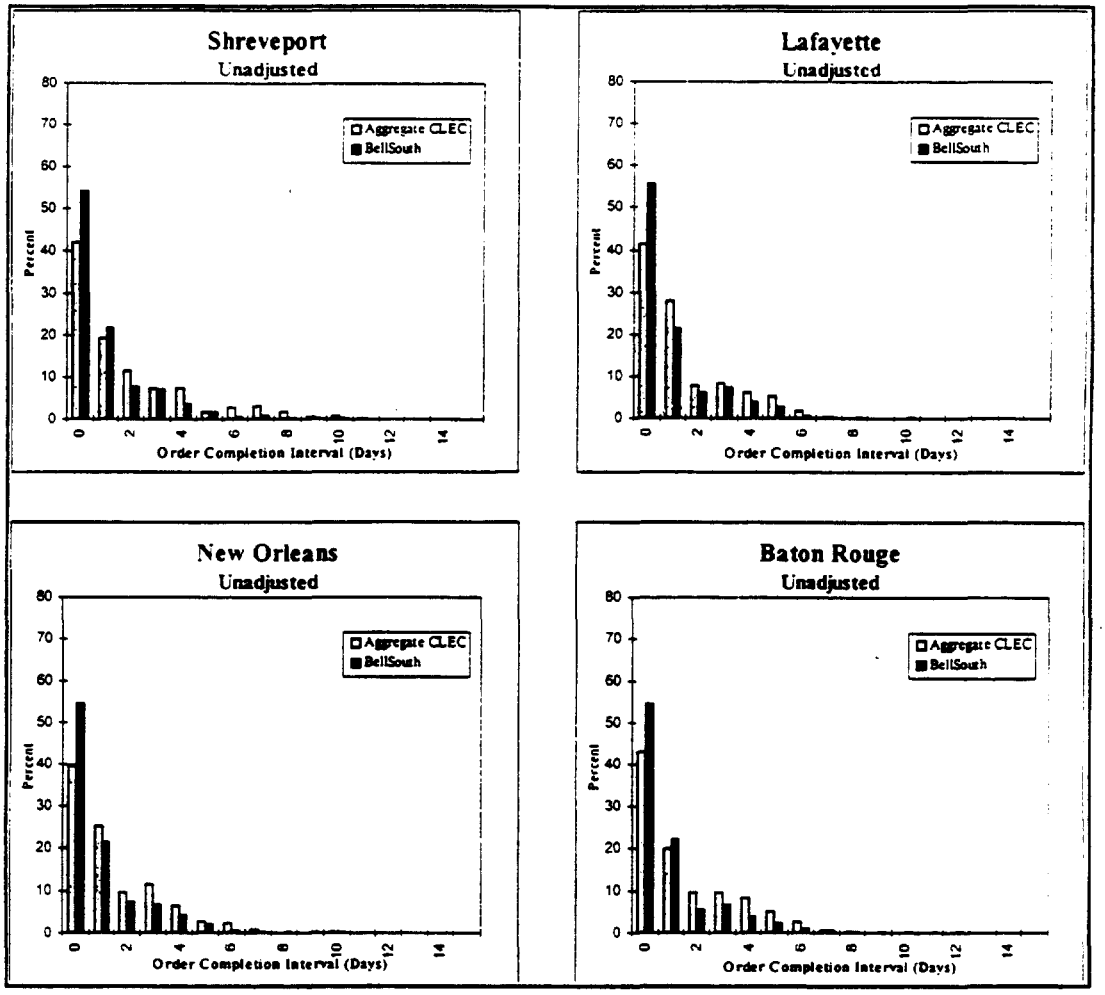
Overall CLEC and BellSouth Relative Frequency Distributions. – We now look at the frequency distributions for each LATA in order to determine whether those distributions are similarly shaped for BellSouth and for the CLECs. Figure 28 shows the four distributions for the unadjusted data, and Figure 29 presents them for the adjusted data.

Figure 28 - Order Completion Interval Frequency Distributions by LATA, Unadjusted Data



These graphs show great similarity, both between the BellSouth and CLECs distributions, and among the four LATA. Notice that in each LATA for the 0 interval (0 - 10 days), the BellSouth bar represents about fifty-five percent of the observations.

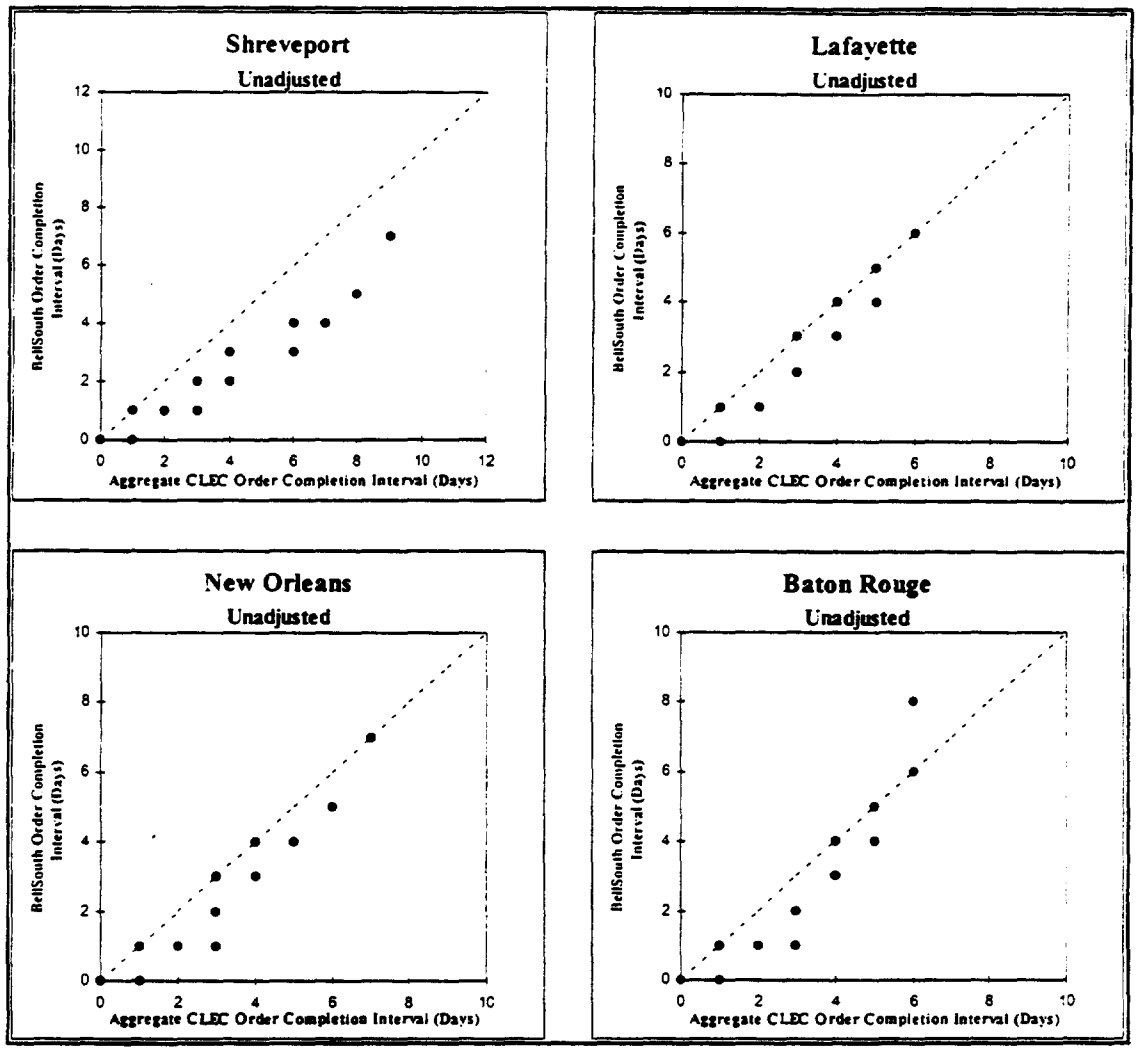
Figure 29 - Order Completion Interval
Frequency Distributions by LATA, Adjusted Data



These distributions for the adjusted data again show much similarity both between BellSouth and CLECs, and among the four LATA. Comparing them to the distributions shown in Figure 28 also shows close similarity. Note that now the 0 interval for BellSouth in each LATA is lower than it was for the unadjusted data.

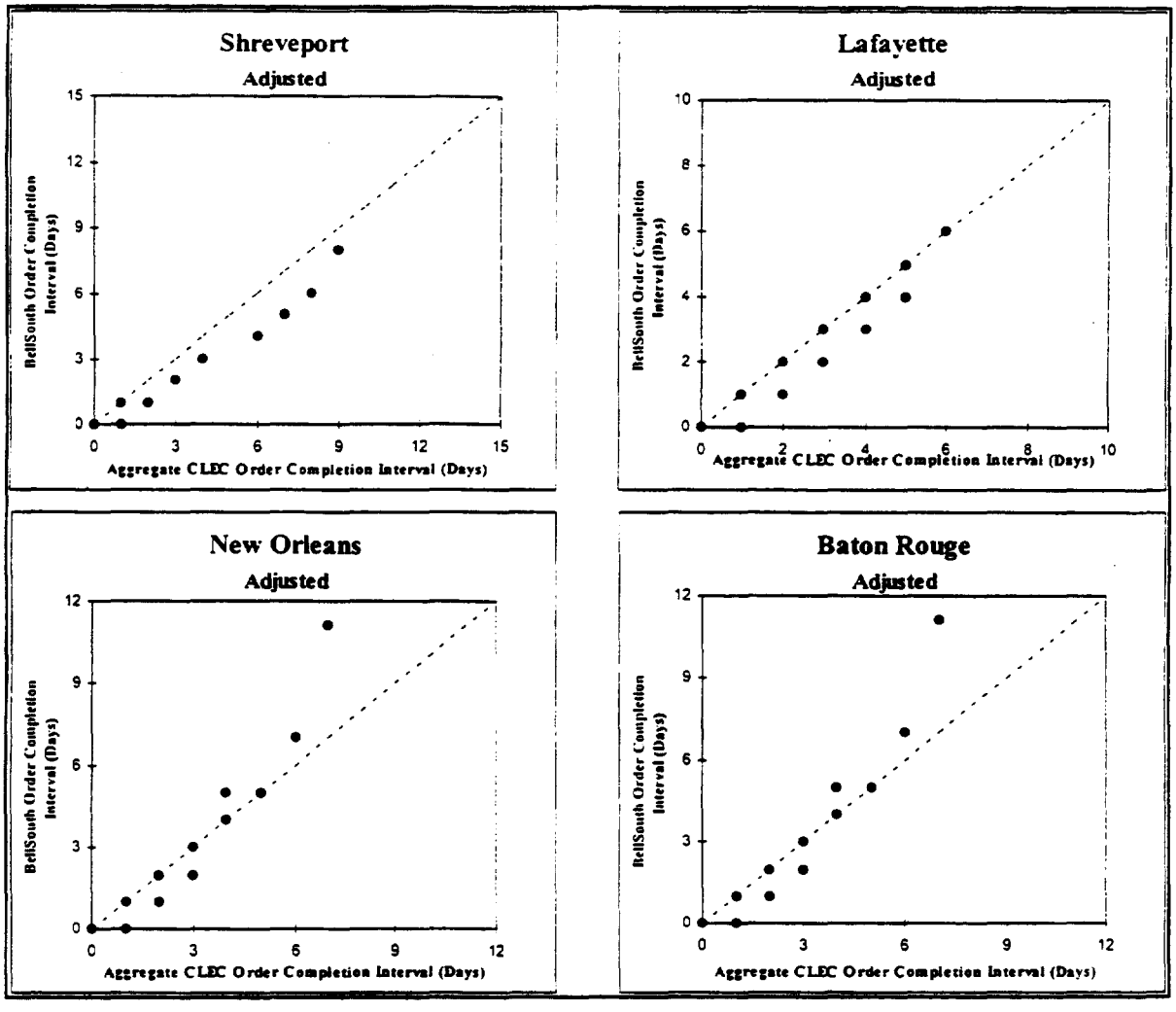
Overall CLEC and BellSouth Quantile Comparisons. – The quantile comparisons for each LATA are presented in Figure 30 for the unadjusted data, and in Figure 31 for the adjusted data.

Figure 30 - Order Completion Interval
Quantile Comparisons by LATA, Unadjusted Data



These quantile comparisons tell a similar story for each LATA, yet we can discern differences among them. The Lafayette comparison is closest to demonstrating equal quantiles for BellSouth and CLECs, and in Baton Rouge we notice that one point (the only point anywhere above the 45-degree line) stands apart by itself.

Figure 31 - Order Completion Interval Quantile Comparisons by LATA, Adjusted Data



The points in the quantile comparisons for the adjusted data are all much closer to falling on the 45-degree line of equality. There is a clear outlier in the New Orleans data after adjustment. Again, the distributions for the four LATA are mostly similar.

Figure 32 – Order Completion Interval
Test Statistics by LATA (August)

Unadjusted	Shreveport		Lafavette		New Orleans		Baton Rouge	
Testing Method	Test Statistic	P-value (percent)	Test Statistic	P-value (percent)	Test Statistic	P-value (percent)	Test Statistic	P-value (percent)
LCUG	-19.87	0.00	-6.26	0.00	-9.09	0.00	-3.45	0.03
FCC	-19.78	0.00	-6.32	0.00	-9.15	0.00	-3.49	0.02
BST	-5.84	0.00	-3.93	0.03	-6.57	0.00	-3.02	0.30

Adjusted	Shreveport		Lafavette		New Orleans		Baton Rouge	
Testing Method	Test Statistic	P-value (percent)	Test Statistic	P-value (percent)	Test Statistic	P-value (percent)	Test Statistic	P-value (percent)
LCUG	-11.44	0.00	-3.99	0.00	2.55	0.54	-2.33	0.98
FCC	-11.44	0.00	-4.03	0.00	2.57	0.51	-2.35	0.93
BST	-4.54	0.00	-1.62	5.79	1.93	3.18	-0.78	22.08

For the unadjusted data, all three tests show significant difference in the means for each of the four LATA. In each case, this difference favors BellSouth. The significance is demonstrated by the fact that each P-value is less than 5 percent. That collection of the difference is demonstrated by the fact that each test statistic is negative.

What is also important to notice is that the magnitudes of the test statistics vary across the LATA. Shreveport has statistics standing at almost 20 (in absolute value), while those statistics for Baton Rouge are about 3.5. What this means is that there is probably some difference between the LATAs, and therefore it is better to separate the data in this way.

The most striking aspect to notice when now looking at the statistics for the adjusted data is how much the test statistics have shrunk in terms of absolute value. The story has not reversed; in Shreveport, Lafayette, and Baton Rouge there is still evidence of a significant difference in favor of BellSouth by the LCUG and FCC tests. The BellSouth proposed test shows a significant difference only in Shrevport. However, in New Orleans all tests now show a difference that is in favor of the CLECs.

For each LATA and for both the unadjusted and the adjusted data, the BellSouth statistic is always lowest in terms of absolute value. What this suggests is that the BellSouth method shows less disparity. If the BellSouth method is based on more accurate assumptions than the LCUG and FCC tests, then there is quite likely less true disparity in the data than at first appears from using the first two tests.

VI. Limitations of Disaggregate Analysis

BellSouth publishes hundreds of Service Quality Measurements (SQMs) every month. Most of these SQMs exist for both BellSouth and the CLECs, although there are some that apply to only one of the two groups. Within the set of SQMs that pertain to both BellSouth and the CLECs, tests of the kind done in this report could be performed. Given all the SQMs and the subsets of service categories that they are reported for, the number of parity tests that can be performed might be very large indeed.

It is important to realize that, due to random fluctuations inherent in statistical testing, BellSouth may fail some tests even though no real service difference actually exist. The chance of this occurring increases with the number of tests that are performed. In order to compensate, an additional level of statistical analysis is needed to interpret the aggregate results of the tests.

In our opinion, the methodology employed should be comprised of both diagnostic graphics, as well as numerical statistical summaries:

- Diagnostic graphics can help one to find oddities in the data. An examination of these oddities will often uncover problems that can be corrected. Examples of diagnostic graphs can be found throughout this document (especially in the appendices).

- Numerical analyses should be used in order determine whether or not failures on some of the tests for differences are statistically significant. For this to be done properly, the methodology used should take into account relationships among the measures.

AT&T has suggested a methodology for use in reviewing the aggregate results of multiple tests, the details of which are provided in Appendix J. The AT&T methodology assumes that all the tests are independent, i.e., observing the outcome of one parity test gives you no information about the outcome of any other.

Many performance measures within the same Service Quality Measurement categories are calculated from a common set of data. While the measures quantify different aspects of performance, the fact that certain common variables are used in the calculations suggests that the measures will not be independent.

A given SQM value (e.g., OSS Response Interval) may also be correlated through time. If you consider the annual business cycle, you will notice that the number of line service requests has an oscillatory nature through the year. Because of this, one might expect many of the performance measures to have similar values in consecutive months. Tests that are based on measures that have month-to-month correlation will also exhibit a correlation.

Appendix J contains a detailed look at what might happen when AT&T's procedure is used on a set of measures that are correlated. Because of concerns about independence we would prefer to use a procedure based on the Bonferroni inequality. The Bonferroni approach is also described in Appendix J.

By using the Bonferroni methodology, we avoid potential problems brought about by a lack of independence in the test statistics. It is also useful to point out that both the Bonferroni methodology and the AT&T proposed methodology are approximately the same when only five tests are aggregated. To be specific, applying AT&T's procedure to five tests, no failures are allowed within a month, and the false alarm rate for each individual test is 1.02 percent. A Bonferroni approach would call for much the same procedure – the individual false alarm rate, though, is exactly 1 percent.

Also, if the number of tests is under ten, then the individual test false alarm rate will be greater than 0.5 percent when a Bonferroni procedure is used. Under conventional testing, the critical values for the individual tests would come from the extreme tail of a theoretical distribution like the standard normal or Student's t distribution (see Glossary). This does not work here. The simulations discussed in Appendix J suggest that the distribution of extreme values may not be modeled well by these distributions.

The quantification of performance is an important aspect of quality management. Therefore it is important that BellSouth continue to measure its performance in many different ways. However, when it comes to making judgements as to whether or not BellSouth is meeting its nondiscriminatory obligation with respect to the service it provides CLECs and their customers, there are potential problems that can arise when the results of too many tests are aggregated. As we have noted, these problems include --

- Dependencies that exist among tests within the same month,
- Dependencies between consecutive monthly measurements, and
- Measures with non-normal distributions.

Therefore we recommend that only the results of five to ten tests be aggregated in any given month for a given geographic level. We also recommend that a methodology based on the Bonferroni inequality be used in the aggregation process.

For monthly testing the three measures studies in this report might be among those considered. They certainly are important enough and appear to be nearly independent. With respect to comparing tests over time, more information is need before we can recommend a procedure. For example, we recommend that data from more months be examined to determine the extent of dependencies across monthly test results.

VII. Interim Conclusions

What has been learned in our work so far? To answer this question, only a brief interim summary has been provided here. We divide our remarks into two parts:

- What we learned, so far, about the alternative statistical methods we were asked to look at and the degree to which a common statistical approach to performance measurement might be taken

- Given our preferred approach, what statistically significant differences, if any, did we see for the three measures examined for Louisiana in August and September of this year

Alternative Statistical Methods. – Methods depend crucially on the underlying data structures and any choice among them cannot be separated from this consideration:

- In the case of the Average OSS Response Interval we had only daily summary averages for BST and for the CLECs. This severely limited our approach to studying statistical significance – so much so that the LCUG and FCC tests could not even be calculated.

- For the Order Completion Interval-Provisioning and Maintenance Average Duration, we were able to compare the LCUG and FCC tests with the BST tests recommended here.

For the most part we found that the LCUG and FCC tests had to make strong assumptions that did not appear warranted in the data we examined. The BST tests, on the other hand, are not subject to such strong assumptions. In summary, the BST approaches appear to work efficiently and can be interpreted as a safe starting point for statistically analyzing differences, if any, between CLEC resale and BST retail customers. This is simply not the case for the LCUG and FCC calculations. Table 1, which follows, provides this summary in tabular format, addressing the specific dimensions quoted as the outset of this report from Louisiana Docket No. U-22252.

Statistical Results of Testing. For two of the three measures, the weight of the evidence gives no strong sign that something other than full parity (or better) exists between BST retail and CLEC resale customers. These measures are Maintenance Average Duration and Average OSS Response Interval:

- For maintenance average duration, the evidence indicates a slight but arguably not statistically significant difference in "favor" of BellSouth for August and a slight but again arguably not a statistically significant difference in "favor" of the CLECs in September.

- For the OSS response interval, the evidence is quite strong that, for August at least, BellSouth could be favoring the CLECs over its own customers.

For the third measure -- order completion interval -- the evidence supports a different conclusion. As the report details, BellSouth appears to be providing service to the CLECs which is statistically significantly slower than it provides to its own retail customers. This difference is not large overall for August. After adjusting for customer mix the difference turns out to be just 0.15 days in August. However, this difference rises to 0.59 days in September, clearly both statistically and operationally significant (and thus warranting further study of the underlying causal structure).

Summary Tables 1 and 2. -- Tables 1 and 2 below provide in tabular form the results we have so far regarding statistical methods (Table 1) and the overall statistical results we found for August and September (Table 2).

In Table 1 it will be evident that, as the Commission surmised, apparent statistically significant differences may, in fact, not be truly significant if the measures are constructed at the wrong level of aggregation. This means that artificial differences may be created because background factors, extraneous to the comparisons of "likes-to-likes," could be influencing one group more than another. A second theme is that both the LCUG and FCC approaches make assumptions we found not to hold; hence those tests cannot be recommended as safe in general. On the other hand, convincing confirmation exists that despite formidable business complexity, we have a statistically principled way of providing statistically sound tests for differences between BST and aggregate CLEC performance measurements.

Table 2 displays the statistically valid testing results we obtained for the three measures studied, including our overall results by month, plus our interpretation of them. Basically the OSS results suggest that there is a statistically significant difference in August but that it "favors" the CLECs and not BellSouth. The Maintenance Average Duration results vary from month-to-month showing no pattern of differential CLEC/BST treatment in either direction. Only for the Order Completion Interval do our results show statistically significant differences that "favor" BellSouth in both months.

More To Come. -- This report is of work in progress. For example, by the time of the November workshop we plan to move from these interim summary comments to the point where we will make recommendations to the Commission. We would like, as noted earlier, to look at the month of October, at least, plus some three-month summaries. These further analyses, among others, will be presented on November 30. At that time we will also present additional analyses based on what we have learned regarding the statistically significant differences found for the order completion interval and its potential causes and implications.

Table 1. – Interim Summary of Required Methods Comparison Made for the Louisiana Commission under Docket U-22252

Testing Proposal	When ILEC & CLEC processes are different and not expected to yield same results	When ILEC actually is employing discriminatory practices.	When assumptions necessary for the statistical test to be valid are not meet
LCUG	Calculating these measures at the level of descriptive reporting required can lead to comparisons that are not "like-to-like." The answer here is not more detail (which pushes against sample size limits) but an analytic summary based on standardized data. That is the approach we have taken. In particular by building upon the CLEC volumes to standardize the BST comparisons, much of this concern can be reduced or avoided.	This test has possible merit and in some settings might even be preferred to that suggested by the FCC, albeit the FCC and LCUG numerical results we saw are virtually identical in most cases and have about the same problems -- notably that the strong assumptions required for validity do not always hold.	For monthly Louisiana results clear evidence exists that the assumptions in the LCUG test fail to hold and, hence this test is invalid for general use. Moreover it cannot be employed at all to statistically study differences in OSS response intervals between BST and the CLECs.
FCC		This measure could work well, if "likes-to-likes" are compared. Required, though, is that strong assumptions hold for it to be valid – something we did not find always to be the case.	This test has the same basic weaknesses as the LCUG approach and is, hence, also unsuitable for general use. Moreover, it makes an additional assumption which does not appear to hold in all settings.
BST		The methods we have recommended will have essentially the same efficiency (or power) as the FCC and LCUG tests to detect differences, should they exist. They are, moreover, completely practical and do not prefer one side over the other.	For individual Louisiana results, possible assumption failures are judged unlikely and no evidence for them was found. For the month-to-month changes more study is needed and this will be covered at the November 30 workshop.

Table 2. – Summary Results of Preferred Testing Approach by Type of Performance Measurement, August and September Separately

Performance Measurement	Difference of "Likes-to-Likes"	BST Test Statistic	Interpretation
Order Completion Interval - Provisioning			For both August and September, the tests done show that statistically significant differences exist favoring BellSouth over the CLECs. For September, moreover, the difference almost certainly are large enough to have operational significance. Both months merit further study and our findings will be given at the November 30th workshop.
August	-0.14 Days	-2.57	
September	-0.59 Days	-8.81	
Maintenance Average Duration			The test statistics for the Maintenance Average Duration are near statistical significance in each month but in opposite directions. No further action seems called for.
August	-1.38 Days	-1.93	
September	2.32 Days	2.43	
OSS Response Time			For OSS Response Time, the test statistics are both positive and for August highly significant, suggesting if anything, that BellSouth is favoring the CLECs over itself.
August	.3197 Seconds	3.78	
September	.1028 Seconds	1.20	

Note: "Statistical Significance" in this report is defined to have been reached when the test statistic is outside the range ± 2. By convention, when the difference is positive, we say the measure suggests that the CLECs resale customers are getting better treatment than BST retail customers. The reverse is true if the sign of the difference is negative. Differences that are +2 or larger are defined therefore to be differences which statistically significantly "favor" the CLECs. Differences that are -2 or smaller are defined to be differences which statistically "favor" BellSouth (see Glossary and Appendix B).

Appendices

A. Credentials and Experience

B. Statistical Calculations for Two Performance Measures -- Completion Interval - Provisioning and Maintenance Average Duration.

C. Order Completion Interval: August Graphics

D. Order Completion Interval: September Graphics

E. Maintenance Average Duration: August Graphics

F. Maintenance Average Duration: September Graphics

G. OSS Average Response Interval

H. LATA: August Graphics

I. LATA: September Graphics

J. Aggregate Assessment of Nondiscrimination - Multiple Test of Parity

K. Glossary of Acronyms and Statistical Terms

Appendix A
Credentials and Experience

Appendix A
Credentials and Experience for Principal Authors

Fritz Scheuren Qualifications

1. I have been a professional mathematical statistician for more than 25 years. I have a BA from Tufts University. My graduate work, both MA and Ph.D., are from The George Washington University, where I continue to teach sampling.

   Since January 1997, I have worked at Ernst & Young LLP as the National Technical Director for Statistics and Statistical Sampling. I now also hold the position of Principal. My diverse experience at Ernst & Young has included managing large audit sampling engagements, designing major inventory sampling efforts, handing disputes with the IRS on statistical matters and providing sampling and statistical advice in many other regulatory settings. My industry experience includes banking and finance, consumer products, healthcare, mining, retail and wholesale trade, and transportation. Much of my recent work here has been on various statistical applications in the telecommunications industry.

   Prior to joining Ernst & Young (from 1994 to 1997), I was a Professor of Statistics at The George Washington University. From 1980 to 1994, I was the Director of the Statistics of Income Division of the Internal Revenue Service (IRS) and, as such, its highest ranking statistician. Prior to joining the IRS, I worked at the Social Security Administration (SSA) where I eventually rose to be its Chief Mathematical Statistician.

2. I am a Fellow and Vice President elect of the American Statistical Association (ASA). In addition, I am currently the Chair of the Scientific and Public Affairs Committee of the ASA. I am also a Fellow of the American Association for the Advancement of Science and a member of the International Statistical Institute (ISI) and the American Society for Quality. Among my many professional roles, I have been the Scientific Secretary of the International Association of Survey Statisticians of ISI and a member of the Committee for Applied and Theoretical Statistics, National Academy of Sciences. In 1995, I received the Shiskin Award for contributions to U.S. economic statistics and in 1998 the Founder's Award, the highest service award given by the ASA.

3. As an internationally known sampling expert, I have published widely on survey design and other statistical problems -- authoring or co-authoring nearly a hundred books, monographs, and papers. Over the years in my statistical practice, I have been a consultant expert, expert witness, and acted as a manager or technical director on all

sampling and statistical aspects of numerous projects, both large and small, for many corporate and government clients. One of my main professional interests has been in developing ways of turning operating data systems into statistical information systems – an area on which I have published extensively. This was particularly important when I was at the IRS and SSA, which have some of the biggest operating data systems in the Federal Government. My large systems experiences were especially relevant to the analyses in this report which had to be developed from BellSouth's truly massive datasets.

Susan Hinkins Qualifications

1. I have been a professional statistician for 20 years. In 1971 I obtained a B.S. in mathematics from the University of Wisconsin-Madison, an M.S. in mathematics in 1973 and a Ph.D. in statistics in 1979 from Montana State University-Bozeman.

   Since July 1998 I have worked at Ernst & Young LLP where I am now Chief Mathematical Statistician for Statistical Sampling. Before coming to Ernst & Young, I was a senior mathematical statistician at the U.S. Internal Revenue Service. My work at the IRS related primarily to business data, notably that on corporations. I was responsible for developing and maintaining a large and complex sample from a population of approximately 4 million corporate returns.

   I have also worked on a large project funded by the Environmental Protection Agency (EPA) to do an exploratory data analysis of a complex sample of all lakes in the U.S., measuring water chemistry and physical characteristics. While working for the EPA, I also coordinated a study to compare various methods for measuring the level of radon and radon-daughters in homes.

2. I am a member of the American Statistical Association (ASA), the Washington Statistical Society, and I am the Secretary/Treasurer of the Montana Chapter of the ASA. I am also a member of the Institute of Mathematical Statistics and the scientific research society, Sigma Xi.

3. My interests and experience have lead me to specialize in the analysis of complex samples, data imputation, and related estimation issues. I have authored and co-authored numerous papers dealing with these issues. Of particular importance in the current context is the work I have done on replicate variance estimation and its application to complex sample data. The replicate approach we recommend in the report to BellSouth grows out of my theoretical work and prior practical applications.

Ed Mulrow Qualifications

1. I have been a professional statistician for more than 10 years. I obtained a BA in mathematics in 1980 from Illinois Wesleyan University, an MS in mathematics from the

University of Utah in 1982 and a Ph.D. in statistics from Colorado State University in 1986.

Since April, 1998, I have worked at Ernst & Young LLP where I am now a manager in the Policy Economics and Quantitative Analysis Group. At Ernst & Young, I have capitalized on my extensive prior defense simulation experience and taken the lead on large scale simulation modeling in commercial business settings. This has included distribution free estimation using normal and near normal data sets.

Before coming to Ernst & Young, I was a senior scientist at Science Applications International Corporation (SAIC) where I was involved in the analyses of current and future defense systems. In addition, I was the project leader for the development of a database system used to track funding for Department of Defense Information Technology projects. I also worked at the National Opinion Research Center (NORC) as a senior sampling statistician, where I developed a prototype sampling system. The system consisted of a data warehouse of all the information needed to sample from several national sampling frames, and software tools that access and process the information. I headed a committee that oversaw the acquisition and use of a geographic information system (GIS), and was the lead statistician for NORC on record linkage projects. Before moving to the defense/business environments, I was an Assistant Professor of Mathematics at Southern Illinois University - Carbondale.

2. I am a member of the American Statistical Association, the Washington Statistical Society, the ASA Statistical Computing and Graphics Section, and the Military Operations Research Society, in addition to managing the membership database for the Caucus for Women in Statistics.

3. I have co-authored statistical articles and refereed papers for several domestic and international journals. My interests and experience lead to special expertise in statistical computing and graphics, time series analysis, record linkage, geographical information systems and the design and development of large databases.

Appendix B
Statistical Calculations for Two Performance Measures – Completion Interval - Provisioning and Maintenance Average Duration

Appendix B
Statistical Calculations for Two Performance Measures: Completion Interval - Provisioning and Maintenance Average Duration

Purpose and Structure

This appendix describes three methods for testing the hypothesis that the CLEC orders are being treated in a comparable manner to the BST orders. Examples are drawn from the Completion Interval - Provisioning measure, but the method also applies to the performance measure Maintenance Average Duration.

First, the model assumptions and properties of the FCC and the LCUG methods are described. Then we describe how the underlying assumptions for these tests are not valid in this situation, and how such model misspecification affects the tests. We describe what we believe is a more reasonable model and our proposed replicate methodology. We provide the formulas for the six test statistics given in the main report, namely the LCUG, the FCC, and the proposed BellSouth method, unadjusted and adjusted. Finally, we summarize the steps for our proposed replicate method, including the data analysis steps and test procedures, and we reiterate the reasons why this method should be adopted.

Basic Theory

Statistical texts generally have at least one section describing the comparisons of two populations, textbooks such as Snedecor and Cochran (1967), Hogg and Craig (1970), and Kempthorne (1973), for example. And often, as in this case, the interest is in comparing the location of the two populations, measured by the mean or the average value. The assumption is often made that the observations are from two normal distributions (the treatment and the control) with the same variance or dispersion but different means. For each population, the observations are assumed to be independent and identically distributed (IID).

These are very strong assumptions and may not hold in many applications. In the performance measures considered up to this time, the underlying distributions are clearly not normal, nor even symmetric distributions. However, the great advantage of considering a comparison of means is that the distribution of the mean value can be approximated by a normal distribution, using the Central Limit Theorem, if the sample sizes are large enough and the underlying distribution is not too skewed. Therefore, a reasonable alternative assumption is that the sample means, say $\bar{x}_1$ and $\bar{x}_2$, are normally distributed and are *independent*. The assumption that the two populations have the same variance is necessary to use the standard test; if the variances are unequal, adjustments must be made to either adjust or approximate a t-distribution for the usual test statistic.

A very important underlying assumption is that the data are the result of a designed experiment, where the "treatments" are assigned randomly to the units of analysis. Any confounding factors or possible blocking effects are taken into account in the design of the experiment and all other assignments are randomized in order to remove bias due to any remaining systematic differences in the units.

For example, in agricultural experiments, location is often considered a blocking effect. Plots that are close together tend to give similar yields due to otherwise uncontrolled effects, such as drainage and fertility gradients. Treatments are assigned at random to plots within each block.

The block effect may be on the mean (fixed effect) or on the variance (random effect), describing correlations between units that are physically close to each other. In this case, we do not have a controlled experiment and this should add an extra note of caution, as emphasized elsewhere.

Consider the simplest general model for the two population comparison. Let x_{1i} denote the performance measurement on BST order i, i=1,...,n_1. Let x_{2j} denote a performance measurement on a CLEC order, j=1,...,n_2. Then the most basic model is

$$x_{1i} = \mu + \varepsilon_i \qquad \text{where } \varepsilon_i \sim \text{IID}(0, \sigma_1^2)$$
$$x_{2j} = \mu + \tau + \delta_j \quad \text{where } \delta_j \sim \text{IID}(0, \sigma_2^2)$$

and the two means $\bar{x}_1$ and $\bar{x}_2$ are independent. If the underlying distributions are not too skewed and the sample size is reasonably large, then one can reasonably approximate the distribution of the difference in the means as normally distributed

$$\bar{x}_1 - \bar{x}_2 \sim N(\tau, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}) \qquad (1)$$

and we are interested in testing whether $\tau = 0$.

FCC Measure. In addition, it can be assumed that the variances are the same in each case, $\sigma_1^2 = \sigma_2^2 = \sigma^2$. That is, it is assumed that the two distributions are the same, except for a possible difference in the means, due to a "treatment" effect.

These are the assumptions used in the FCC measure. A pooled estimate of the variance is used, s_p^2, and the resulting t-test is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{1/n_1 + 1/n_2}}$$

with $n_1 + n_2 - 2$ degrees of freedom. It often turns out to be the case that the sample sizes will be large enough so that the normal, or Z, distribution can be used rather than the t-distribution.

In at least some cases in the Louisiana data that we have studied, it does not appear that the assumption of equal variance is valid. There are two other measures that are being

considered - the LCUG and the measure that we prefer. Neither of these measures assumes equal variance.

LCUG measure. Rather than assume that the variances are equal, the LCUG estimate simply uses the BST population variance as the standard for comparison. The t-test then has n_1-1 degrees of freedom and the test statistic is of the form

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_1\sqrt{1/n_1 + 1/n_2}}.$$

Again, if the assumptions for the test hold, the BST sample size is usually sufficiently large that the normal distribution is appropriate for calculating p-values.

If the two distributions are identical except for location, then the FCC is a test of the equality of the two distributions. If the variances are not equal, then the interpretation of the test is endangered. If one is concerned about the assumption that the variances are equal, then using the BST variance is a reasonable alternative.

Even if the variances in fact are equal, it costs very little to use the BST variance rather than the pooled variance. And if the number of BST cases is much greater than the number of CLEC cases, it could even be preferred because of concerns about pooling the data with relatively few CLEC cases.
If the variances are unequal, then the correct test would be based on equation (1) and the test would be of the form

$$t' = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}.$$

If in fact the BST variance is less than the CLEC variance, then the LCUG test is more stringent, harder to "pass" than the correct test. If the BST variance is greater than the CLEC variance, then the LCUG test is not as stringent as the test using both variances, as in equation (1). Our test, which will be described in this appendix, also does not assume equal variances, and if the assumption of independence holds, our test uses a "correct" variance estimate in that it estimates the variance in (1).

First Steps in Data Analysis

The first performance measure that we analyzed was the Completion Interval-Provisioning for the months August and September. This is measured in days and estimates are made separately for dispatched and non-dispatched orders, and also separately by the type of order: "residence," "business," or "special designed" orders, and by two classes determined by the number of circuits.

The "Non-Designed" Maintenance Average Duration performance measure is measured in hours, and estimates are made separately for dispatched and non-dispatched orders, and separately for "residence" and "business" orders. The examples used in this discussion come from the Completion Interval - Provisioning measure, but the techniques apply to both measures.

The first step in the data analysis was to verify the data set. This was done by calculating the estimates and comparing them to the published estimates on the BST internet website (https://clec.bellsouth.com).

Trimming. The underlying distribution of the orders is clearly not normal, but rather skewed with a very long upper-tail. (See Appendices C and D.) Extreme data values may be correct, but since they are rare measurements, they may be considered to be statistical outliers. Or they may be values that should not be in the analysis data set because of errors in the measurement or in selecting the data.

The arithmetic average is extremely sensitive to outliers; a single large value, possibly an erroneous value, can significantly distort the mean value. And by inflating the error variance, this also affects conclusions about whether $\tau = 0$. A useful technique, coming from the field of robust statistical analysis – for example Huber (1981), or Wiens, Wu, Zhou, (1998) -- is to trim a very small proportion from the tails of the distribution before calculating the means. The resulting mean is referred to as a trimmed mean. Trimming is beneficial in that it speeds the convergence of the distribution of the means to a normal distribution. Only extreme values are trimmed, and in many cases the data being trimmed are, in fact, data that might not be used in the analysis on other grounds.

In the first analysis of the verified Completion Interval-Provisioning measure, after removing data that were clearly in error or were not applicable, we looked at the cases that represented the largest 0.01% of the BST distribution. In the August data, this corresponded to orders with completion intervals greater than 99 days. All of these were BellSouth orders.

In examining the largest 11 individual examples that would be removed from analysis, we found that only 1 of the 11 cases was a valid case where the completion interval was unusually large. The other 10 cases were examples of cases that should not have been included in the analysis.

Of the 11 largest values, eight were orders which are "official BellSouth orders"; these are internal jobs which are not real orders but which needed an order number for tracking purposes. These orders can be identified using the data field "general class service" and such orders were subsequently removed from the analysis data file.

Two of the cases were orders where the customer requested a later due date than offered by BellSouth. The customer called in February to place an order for August, for example. There is no easy way to identify such cases in general, in order to remove them from analysis.[1] The system is not yet stable; hence, there may be other types of data points that should not be included or that are not measured correctly. A very slight trimming is needed in order to put the central limit theorem argument on firm ground.

[1] As a result of our analysis, we eliminated further records from data analysis, both above and below the 99 days, using the information regarding general class of service (official BellSouth orders). The subsequent trimming only removed 15 BST cases from the August BST file and 13 BST cases in September.

We now have a data file of CLEC orders to compare with a data file of BST orders. However, both the tests described earlier treat the problem as if the observations come from a designed study where treatments are assigned at random to units in the population. This is not the case here; rather what we have in the BST and CLEC comparison is an example of an observational study. This is an extremely important distinction that cannot be ignored.

Observational Studies

As is well known, randomization in a designed study is a very powerful tool in removing or reducing bias due to systematic differences in units. A few of the references dealing with the importance of randomization and the difficulties inherent in observational studies would include Fisher (1925); Cochran and Rubin (1973), Holland (1986), Rosenbaum (1987).

In an observational study such as this, there may be variables other than the "treatment" that affect the dependent variable (performance measure) and these variables may be differently distributed across the treatment groups. With the presence of confounding variables, a basic approach would be to list the major confounding variables and find some method of removing or reducing the biases that they may cause.

It is necessary to consider the business structure. Like the agricultural example, "location" in the business should be considered for blocking effects. It seems reasonable that there may be a positive correlation between performance measures within a business unit or a geographic location. The use of the "wire center" was considered as the best location measure. Scatterplots are presented in the main report that illustrate that there is a correlation between BST and CLEC measures.

Blocking or clustering effects in the data mean that the observations are neither independent nor identically distributed, two assumptions made in the LCUG and FCC testing approaches. A positive correlation between the performance for orders within a location would mean that the variance estimates used in both the FCC and the LCUG tests are biased and, in particular, they underestimate the variability in the differences.

Additionally, one might expect that the time of the order may be correlated with the performance; clearly extremes in weather would affect the performance. And one might certainly expect a time and location interaction effect. In the BellSouth comparisons, the data are examined on a monthly basis, which is determined by when the order is completed. Weather conditions occur on a shorter time frame. In the case of these two performance measures, each month is divided into just two components, the first half of the month and the last half. These divisions are made so that the time is divided up as evenly as possible by the days of the week as well.

In addition, for a given performance measure, there may be different types of orders and different types of customers. For example, in the provisioning example, the measurements are compared by dispatch vs non-dispatch, residence vs business vs "special designed", and by the number of circuits. In addition one might want to consider the type of order in terms

of "new" vs "change" vs "transfer". It appears, for instance, that a "new" order takes noticeably longer to finish than a "change" or "transfer."

Finally, if one were designing a study to compare the CLEC to the BST "treatment," one would make sure that the same number of CLEC and BST cases were assigned by the location, by time, and by the type of order. By using random assignment to assign a population unit as either a CLEC or a BST, one would be protected against the possibility of other unsuspected sources of bias. That is, if there is another variable that affects the performance measure, by using random assignment one is likely to assign approximately the same proportion of BST and CLEC orders across the distribution of this variable.

Without random assignment, there is the possibility that the distribution of these confounding variables is very different for the BST orders than for the CLEC. For example, if "new" service tends to take longer than the other service types and one month 50% of the CLEC orders are "new" compared to 25% of the BST orders, then the simple comparison will be biased. The bias may work in either direction, depending on the distribution of the observed data. In the example above, the simple estimate would overestimate the difference between the BST and the CLEC performance, making the CLEC customer performance look worse than that for BST customers since CLEC provisioning would appear to take longer. If the distribution had been out of balance in the other direction, with a higher percentage of new BST orders than new CLEC orders, then the simple estimate would have made the CLEC performance look better than it was.

In summary, the assumptions made for both the FCC and the LCUG tests are not valid. The observations are not likely to be independent and identically distributed. Assumption failures may affect both the numerator (the point estimate of the difference) and the denominator (the estimate of its variability). Clustering effects in the data, resulting in a positive correlation between observations in the same wire center, would mean that the variance estimates used in both the FCC and the LCUG measures are biased. And, in particular, they will underestimate the variability in the differences. In addition, effects due to time or order type may bias the estimate of difference.

Adjusted Estimates. In an observational study, bias is a major concern. There are many references for estimation techniques using data from observational studies. There are two principal strategies for reducing bias in observational studies (Cochran and Rubin, 1973): matching and model related adjustments. When the confounding variables are classification measurements, as they are in this case (new vs. change, time 1 vs. time 2 etc), then both matching and model based strategies lead essentially to the same simple adjustment.

Suppose there are j=1,...,J classes defined by the confounding variables. (One class might be new service in a residence, dispatched service, with less than 10 circuits, finished in time period 1, in wire center "a.") Suppose there are n_{2j} CLEC cases and n_{1j} BST cases in class j with $n_{2j} \geq 0$. The following

estimate of the difference in the means will be subject only to residual biases due to confounding variables "missed" in the classification (Cochran and Rubin, 1973):

$$\hat{D} = \frac{\sum_j n_{2j}(\bar{x}_{1j} - \bar{x}_{2j})}{n_2} \qquad (3)$$

where n_2 is the total number of CLEC observations. Note that there may be classes for which there are BST units but no CLEC units. If this occurs, these BST units are not used in the comparison. This is reasonable when comparing "likes to likes," as required by the Louisiana Commission. Data unique to the BST process should not be used in such a comparison. It is very unlikely that there will ever be a case where there are CLEC observations in a class but no BST observations. So this concern is not directly addressed here; we simply have not seen any examples. In other settings, though, there may be no retail analogue for certain resale activities. Cases with no retail analogue are out of scope in this analysis.

The estimate in equation (3) can also be written as a difference between an adjusted BST mean and the CLEC mean, where BST cases have been weighted or adjusted to represent the CLEC distribution by class. That is,

$$\hat{D} = \bar{x}_{1A} - \bar{x}_2 \qquad (4)$$

where $\bar{x}_{1A}$ is the ILEC adjusted mean:

$$\bar{x}_{1A} = \frac{\sum_j \sum_{i=1}^{n_{1j}} w_j x_{1ji}}{\sum_j \sum_{i=1}^{n_{1j}} w_j}$$

where the weight for BST cases in class j is $w_j = n_{2j}/n_{1j}$, the number of CLEC cases in class j divided by the number of BST cases in class j. The sum of the weights is then simply n_2. The weights adjust the BST cases so that they are "like" the CLEC cases in number and distribution among classes. This is referred to as the adjusted mean or the adjusted estimate.

If in fact we have included all important factors, then $\hat{D}$ is an unbiased estimator for the difference in means. Notice that this estimate can be "rolled up" (or down) to provide reasonable estimates at various levels of aggregation.

An Example. The simple example from Section 3 will be used to illustrate how the adjustments are calculated. In this example, we have the following number of orders:

Service Provider	New Orders	Change Orders
Provider A	n_{11}=30	n_{12}=90
Provider B	n_{21}=60	n_{22}=30

There are only two classes, j=1,2. Recall that in this example there is no discrepancy in the means, by class. For each

provider, the mean is 2 days for class j=1, new orders, and the mean is 1 day for class 2, change orders.

Suppose we want to adjust provider A's distribution to compare to provider B. Then in the notation used in this appendix, we have

$$n_{11}=30,\ n_{12}=90,\ n_1=120$$

$$n_{21}=60,\ n_{22}=30,\ n_2=90$$

Using equation (3), the estimate of the difference would be

$$\hat{D} = \frac{60*(2-2)+30*(1-1)}{90} = 0.$$

The unadjusted means are 1.25 for provider A and 1.67 for provider B. The adjusted mean for provider A would be calculated using weights $w_j = n_{2j}/n_{1j}$, or in this case

$$w_1 = 60/30 = 2$$
$$w_2 = 30/90 = 1/3$$

and the adjusted mean for provider A would be

$$\bar{x}_{1A} = \frac{2*30*2+\frac{1}{3}*90*1}{2*30+90/3} = 1.67.$$

Because there was no discrepancy in the means, by class, the adjusted mean for provider A is equal to the mean for provider B.

Replicate Variance Estimation

The estimate $\hat{D}$ from equation (3) or (4) then is a better estimate of the difference between the mean performance for the BST orders and the mean performance for the CLEC orders. We now need a variance estimate for $\hat{D}$.

Replicate variance estimation can result in a nearly unbiased estimate of the variance for complex data structures like those which exist with the BellSouth data. A description of the basic technique can be found in Wolter (1985). The basic idea is to randomly divide the given sample into G groups, where each group has approximately the same number of wire centers. In each group g, calculate an estimate of the parameter of interest, say $\bar{d}_g$. Let $\bar{\bar{d}}$ be the average of the replicate means $\bar{d}_g$. Then the replicate variance estimate of $\bar{\bar{d}}$ is

$$v_1 = \mathrm{Var}(\bar{\bar{d}}) = \frac{1}{G}\frac{1}{(G-1)}\sum_g (\bar{d}_g - \bar{\bar{d}})^2 \quad . \quad (5)$$

In our problem, however, the estimate we are interested in is $\hat{D}$ which is not generally equal to $\bar{\bar{d}}$. We can use v_1 as an estimate of $\hat{D}$ or the alternative estimator

$$v_2 = \mathrm{Var}(\hat{D}) = \frac{1}{G}\frac{1}{(G-1)}\sum_g (\bar{d}_g - \hat{D})^2 \quad . \quad (6)$$

We have chosen to use expression (6) for the calculations of test statistics employed in the main report and in the four appendices C-F.

Detailed Problem Formulation

In what follows, an explicit attempt is made to describe the specific estimation procedure we recommend for Louisiana that compares "like-to-like" and that captures variances adequately. We are concerned about dependences which could exist in service based on where the customer is geographically or when the transaction occurs. Protecting against this possibility is one of the main motivations for our approach. Ease and simplicity are others.

In all cases, we will want to consider the following in constructing our estimates:

Wire Centers - There are approximately 228 wire centers[2] managed by BellSouth in its four LATA in Louisiana: New Orleans (67), Baton Rouge (31), Lafayette (42), and Shreveport (88)

Time - Service varies over time for many reasons, weather being perhaps the most important. To deal with this source of variation, each month's data will be divided into two approximately equal halves. Weekly increments might be better but could be too fine-grained and are inconvenient since the reporting is monthly and not even in four week periods (which arguably would be better).

Other Factors - There may be other factors considered important in their effect on performance, such as the order type in the Completion Interval-Provisioning. These have to be accounted for too.

Individual Service Order - Lastly, we have the individual order itself

Replicate Construction. We want to define the replicates only once. The replicates were defined, as described here, using the August Completion Interval-Provisioning measure and the resulting definition of the replicates by wire center was used for both performance measures in all time periods.

The wire centers were sorted within LATA by the total CLEC activity, in terms of the number of orders. Wire centers with no CLEC activity in this month were also included, with zero

[2] In the preliminary data analysis, there were 228 wire centers. But because the mapping of phone numbers to wire centers was not complete, there were phone numbers that could not be matched to one of these wire centers. These numbers were mapped into four "dummy" wire centers according to the area code of the phone number. The resulting wire centers were not assigned to a LATA but were instead put into a "missing" category. New Orleans LATA corresponds to LATA 490, Baton Rouge is LATA 492, Lafayette corresponds to 488 and Shreveport corresponds to 486.

activity. The LATA were ordered and the wire centers were ordered within LATA. Within the first LATA, the wire centers were ordered from largest to smallest. In the next LATA, the wire centers were ordered from smallest to largest, etc. We then systematically divided the 232 wire centers into 30 roughly equal groups (of about 7 wire centers). This was done by taking the ordered list and splitting it into "zones" of 30 wire centers each, randomly assigning a wire center to a group until all were assigned, then repeating the process independently for the next zone of 30 wire centers, and so on until all had been assigned.

Estimator Construction. The estimator $\hat{D}$ is calculated as in equation (3), using classes defined by wire center and time at least. The replicates are assigned, by wire center. The adjusted replicate estimates $\bar{d}_{Ag}$, g= 1,.. 30, are calculated using equation (3) but summing only over the cases in the wire centers defined to be in replicate g.

These $\bar{d}_{Ag}$ are identically distributed by construction and independent by randomization. If there is a lot of CLEC activity, they may also be approximately normally distributed. Using the replicate structure we estimate the variance for the adjusted estimate as

$$s_{rA}^2 = \frac{1}{29}\sum_{g=1}^{30}\left(\bar{d}_{Ag} - \hat{D}\right)^2$$

and the resulting statistic

$$t = \frac{\hat{D}}{s_{rA}/\sqrt{30}}$$

is compared to the Student's t-distribution with 29 degrees of freedom, as the reference distribution, for calculating p-values. The p-values are the probability of seeing a value as extreme or more extreme then the observed value of t. That is, if t is positive, the probability of a value greater than or equal to t is calculated, using the Student's t with 29 degrees of freedom as the reference distribution. If t is negative, the probability of a value less than or equal to the observed t is calculated.
Using the replicate variance estimate applied to the adjusted estimate of the difference protects against model misspecification. This test does not rely on the assumption that the data are IID and it corrects for bias due to the structure of the data. Using this method, a confidence interval can be constructed for the difference in the means. A reasonable interval is the 95% confidence interval. Using a Z-test, the multiplier is 1.96 which is often rounded up to 2.00. Using a t-distribution with 29 degrees of freedom, the coefficient is 2.045. For all practical purposes, these are equivalent. There is no loss in power in adopting the replicate measure over the FCC or the LCUG measure.

The Six Test Statistics Compared in the Main Report

The test statistic described in the previous section is the method we propose for the comparisons, and, in the main report, it is referred to as the BellSouth test for adjusted data. It adjusts the BellSouth data to make it more similar in

distribution to the observed CLEC data, and it uses a replicate variance estimator.

For comparison purposes, we can also calculate a replicate estimator for unadjusted data and we can calculate the LCUG measure and the FCC measure using adjusted BellSouth data.

The replicate variance estimate for the unadjusted data would be calculated using replicate means $\bar{d}_g = \bar{x}_{1g} - \bar{x}_{2g}$, the difference between the simple means of the BellSouth and the CLEC data in replicate g:. Replicates are only used if there are CLEC data. The associated replicate estimate of the variance for the unadjusted data is

$$s_r^2 = \frac{1}{G-1}\sum_{g=1}^{G}\left(\bar{d}_g - (\bar{x}_1 - \bar{x}_2)\right)^2$$

where there are G replicates.

For the LCUG and FCC statistics applied to the adjusted data, a weighted s^2 is calculated for the BellSouth data as

$$s_{1A}^2 = \frac{\sum_j \sum_{i=1}^{n_{1j}} w_j (x_{1i} - \bar{x}_{1A})^2}{\sum_{j,i} w_j - 1}.$$

Recalling that the sum of the weights is n_2, in this case, the adjusted pooled variance for the FCC test is then

$$s_{pA}^2 = \frac{(n_2 - 1)(s_{1A}^2 + s_2^2)}{2n_2 - 2}.$$

Using the notation developed here, the tests shown in the main report are calculated as follows, where G indicates the total number of replicates used.

Summary of Calculations.

	Unadjusted Data	**Adjusted Data**
LCUG Test	$\frac{\bar{x}_1 - \bar{x}_2}{s_1\sqrt{1/n_1 + 1/n_2}}$	$\frac{\bar{x}_{1A} - \bar{x}_2}{s_{1A}\sqrt{2/n_2}}$
FCC Test	$\frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{1/n_1 + 1/n_2}}$	$\frac{\bar{x}_{1A} - \bar{x}_2}{s_{pA}\sqrt{2/n_2}}$
BellSouth	$\frac{\bar{x}_1 - \bar{x}_2}{s_r/\sqrt{G}}$	$\frac{\bar{x}_{1A} - \bar{x}_2}{s_{rA}/\sqrt{G}}$

Performance Measured as a Proportion

If the performance measure is a proportion or a percentage of cases which possess some characteristic, such as the proportion of orders taking less than two days to finish, then these methods also apply. It may not be immediately obvious, but proportions can be placed in the same framework as sample means.

A proportion can be calculated by measuring a variable x_i for each case, where x_i=1 if the unit has the characteristic of interest (less than 2 days to complete, for example) and x_i=0 if the unit does not have the characteristic of interest. If we have n cases, then the proportion p of orders with the characteristic of interest is calculated as the mean of the x values, $\bar{x}$.

In this way, the tests can be formulated for proportions using the equations given in this appendix. For example, the sample means within classes become p_{1j} and p_{2j}, the proportion of BellSouth orders and CLEC orders, respectively, in class j. The adjusted estimate of the difference is then

$$\hat{D} = \sum_j n_{2j}(p_{1j} - p_{2j}) \Big/ n_2$$

Outline for the Proposed Replicate Data Analysis

The proposed BellSouth procedure is the replicate method applied to the adjusted data. The steps in the data analysis and test calculation that we propose can be summarized as follows:

1. Verify that we have the correct data set, by comparing to the published estimates on the BST internet website (https://clec.bellsouth.com).
2. Remove any additional data values that are not pertinent to analysis (official BellSouth orders for example)
3. If necessary, trim a very small proportion from the tail(s) of the distribution. (In some cases, the original BellSouth data procedure already included an upper or lower bound on data to be used for analysis.)
4. Put the replicate indicator on the data file and define the time classification.
5. Determine if there are other important classifications that should be used as well, such as order type.
6. For every class defined in steps 4 and 5, calculate the difference $d_j = \bar{x}_{1j} - \bar{x}_{2j}$. In one pass through the data files, a file can be built containing n_{2j}, n_{1j}, and d_j for all classes j.
7. From this data file, estimates of the difference in means and t-tests to test the hypothesis of nondiscriminatory treatment can be calculated for any level of aggregation at the LATA level and above.

Conclusions

The proposed replicate methodology compares "like to like" and it protects against failure of the assumptions of independence. The BellSouth procedure compares "like to like" by adjusting the BST distribution to be more similar to the CLEC distribution. It is not fair to compare CLEC results to BST orders that are intrinsically different. The bias due to the fact that the data come from an observational study makes a difference.

By respecting the business structure and using replicate variance estimates, the BellSouth procedure requires very few assumptions about the underlying distribution. In particular, it does not require the assumption that the observations are IID. In the Completion Order Provisioning examples in the main report, we saw that the adjustments and the use of the replicate variance estimate made a noticeable difference in the results. Not using the adjusted replicate method would have lead to very misleading results.

Insurance against model misspecification costs very little in this case. When the assumptions hold, there is a minimal loss in power using the replicate method compared to the FCC or LCUG method (2.04 vs 2.00 for the 5% two-sided significance level.) This is a small price to pay for a measure of protection against bias due to model misspecification. In addition, this procedure is of computationally modest cost to do routinely and it provides much flexibility in computing estimates and tests.

In conclusion, for these two measures and for other measures like them, the BellSouth adjusted replicate procedure should be highly successful and should be adopted. For a small price, it offers insurance against failure of the assumptions. And when the FCC and LCUG assumptions do hold, this method works just as well as they do. Even if a statistically significant difference is found, however, observational studies cannot assign cause. That is. a statistically significant difference in an observational study does not lead to a conclusion regarding discrimination without additional information.

References

Cochran, W.G. and Rubin, D.B. (1973), Controlling bias in observational studies: a review, *Sankhya A* , **35**, 417-416.

Fisher, R.A. (1925), *Statistical Methods for Research Workers*, Oliver and Boyd.

Holland P.W. (1986), Statistics and Causal Inference, with Discussion, *Journal of the American Statistical Association*, **81**, 945-970.

Hogg, R.V. and Craig, A.T. (1970), *Introduction to Mathematical Statistics*, Macmillan Company, New York.

Kempthorne, O. (1973), *The Design and Analysis of Experiments*, Robert E. Krieger Publishing company, New York

Rosenbaum, P. (1987), The Role of a Second Control Group in an Observational Study, *Statistical Science*, **2**, 292-316

Snedecor, G. and Cochran, W. (1967), *Statistical Methods*, Iowa State University Press, Ames, Iowa.

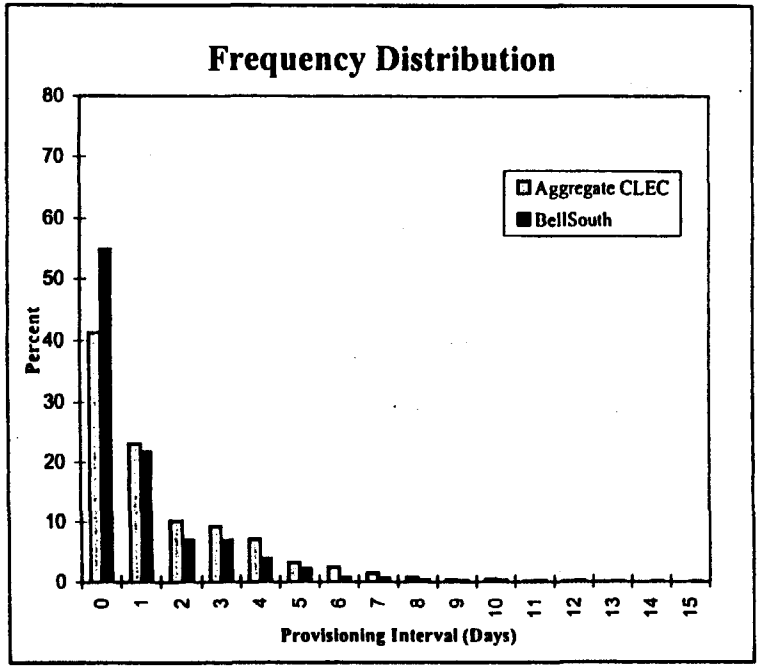
Wiens, D.P., Wu, E.K.H, and Zhou, J. (1998), On the trimmed mean and minimax-variance L-estimation in Kolmogorov neighborhoods, *The Canadian Journal of Statistics*, **26**, 231-238.

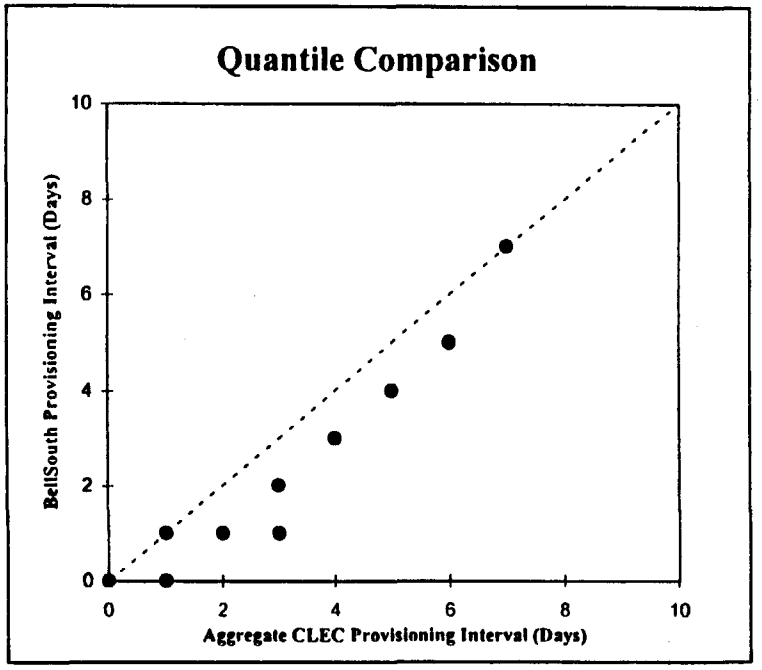
Wolter, K. (1985), *Introduction to Variance Estimation*, Springer-Verlag, New York.s

Appendix C
Order Completion Interval (OCI) - August Graphics

I. Graphical Representations

Unadjusted
August BellSouth and CLEC Completion Interval-Provisioning
All Cases





Descriptive Measures

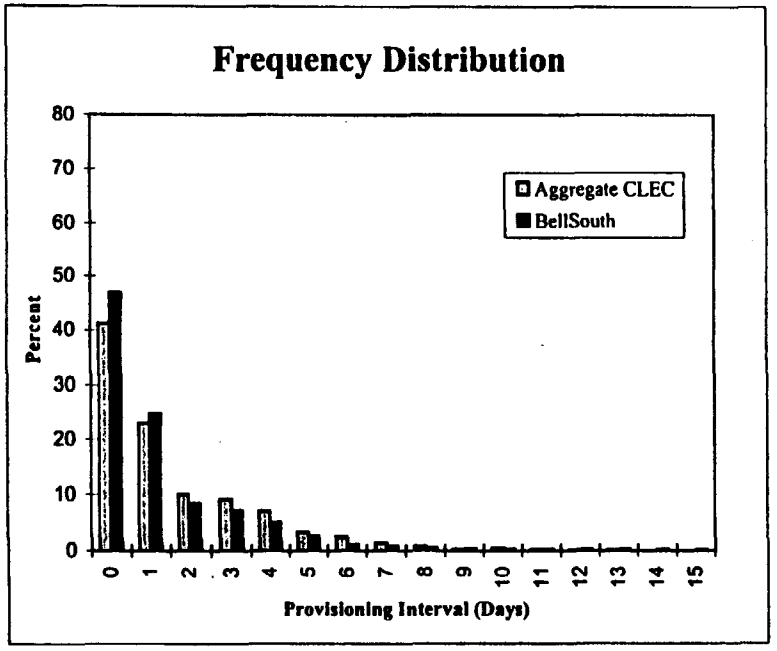
Service Provider	Mean	Standard Deviation
BST	1.20	2.78
CLEC	1.62	2.26
Difference	-0.42	

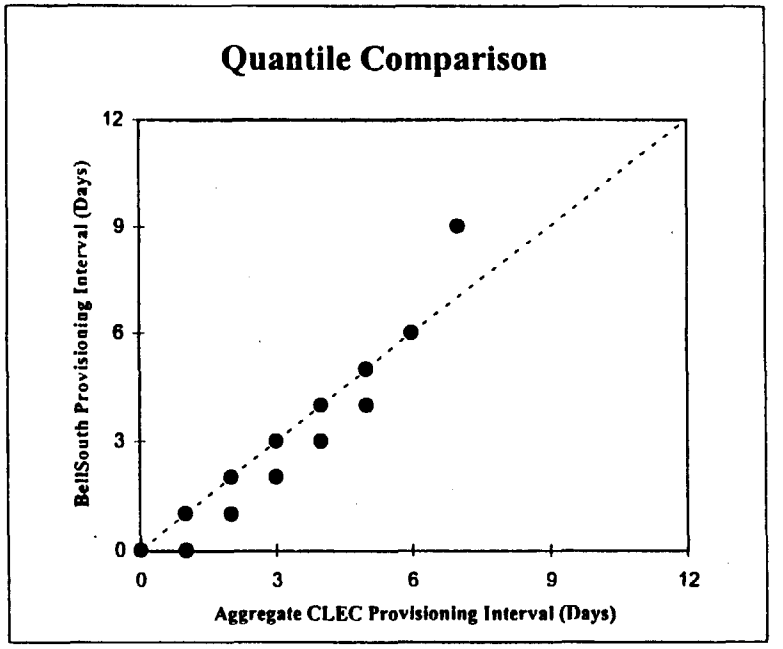
Analytic Measures

Testing Method	Test Statistic	P-value (percent)
LCUG	-18.70	0.0000
FCC	-18.83	0.0000
BST	-9.02	0.0000

Data used in analysis does not include any records with missed appointments due to customer rescheduling or records corresponding to official services.
The application of statistical trimming removed records with completion interval-provisioning of above 99 days. This resulted in the removal of no CLEC records and 0.004% of the BellSouth records.

Adjusted
August BellSouth and CLEC Completion Interval-Provisioning
All Cases





Descriptive Measures

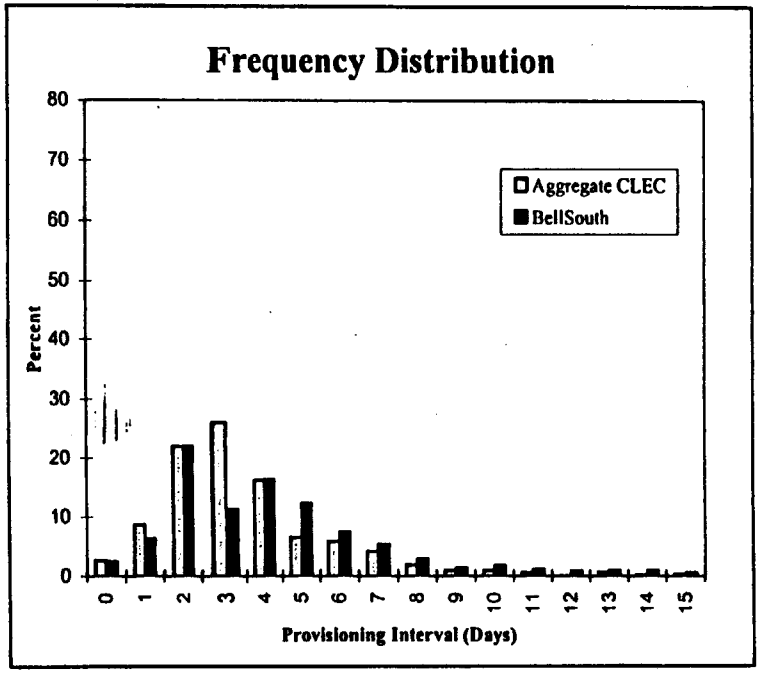
Service Provider	Mean	Standard Deviation
BST	1.48	2.95
CLEC	1.62	2.26
Difference	-0.14	

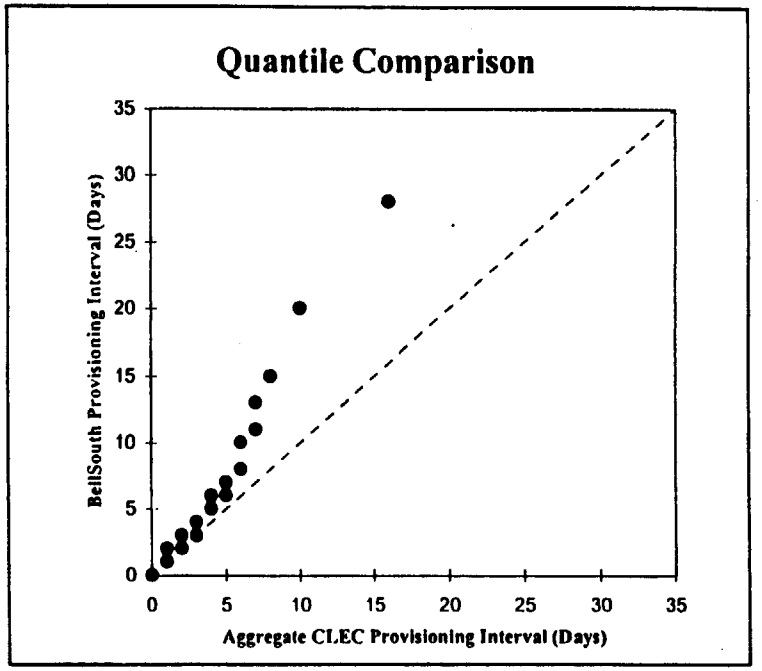
Analytic Measures

Testing Method	Test Statistic	P-value (percent)
LCUG	-6.08	0.0000
FCC	-6.13	0.0000
BST	-2.57	0.7774

Data used in analysis does not include any records with missed appointments due to customer rescheduling or records corresponding to official services. The application of statistical trimming removed records with completion interval-provisioning of above 99 days. This resulted in the removal of no CLEC records and 0.004% of the BellSouth records.

Unadjusted
August BellSouth and CLEC Completion Interval-Provisioning
Dispatched Cases





Descriptive Measures

Service Provider	Mean	Standard Deviation
BST	5.70	7.14
CLEC	3.99	3.77
Difference	1.71	

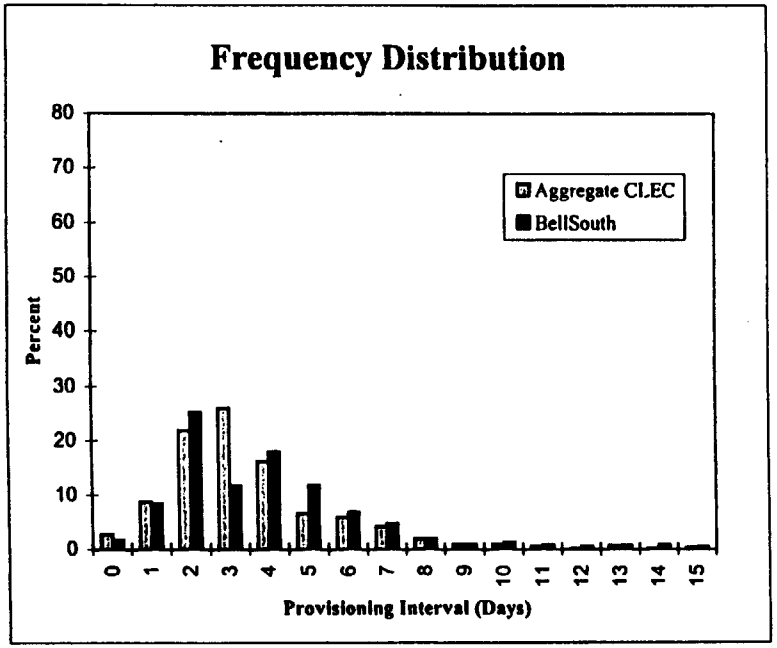
Analytic Measures

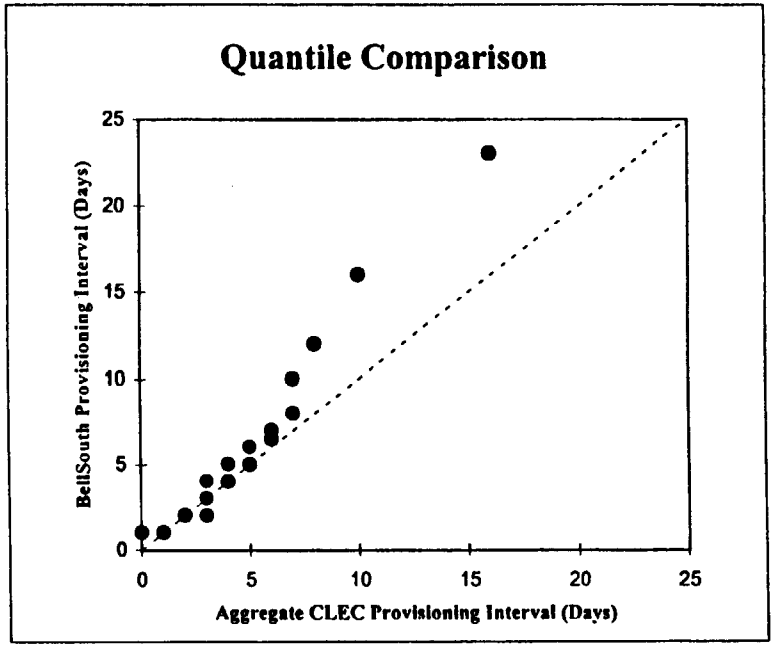
Testing Method	Test Statistic	P-value (percent)
LCUG	8.37	0.0000
FCC	8.53	0.0000
BST	7.13	0.0000

Data used in analysis does not include any records with missed appointments due to customer rescheduling or records corresponding to official services.

The application of statistical trimming removed records with completion interval-provisioning of above 99 days. This resulted in the removal of no CLEC records and 0.004% of the BellSouth records.

Adjusted
August BellSouth and CLEC Completion Interval-Provisioning
Dispatched Cases





Descriptive Measures

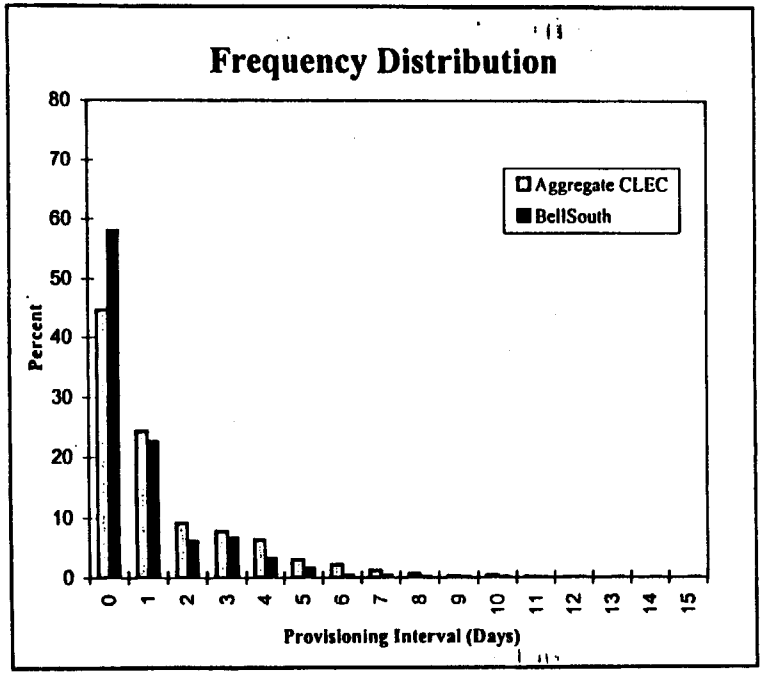
Service Provider	Mean	Standard Deviation
BST	4.88	5.84
CLEC	3.99	3.77
Difference	0.89	

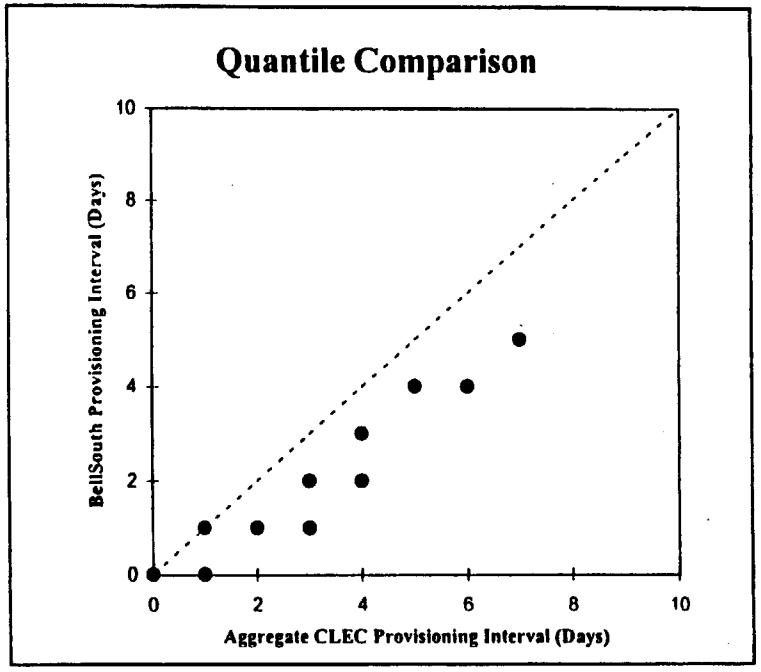
Analytic Measures

Testing Method	Test Statistic	P-value (percent)
LCUG	5.34	0.0000
FCC	5.42	0.0000
BST	6.41	0.0000

Data used in analysis does not include any records with missed appointments due to customer rescheduling or records corresponding to official services. The application of statistical trimming removed records with completion interval-provisioning of above 99 days. This resulted in the removal of no CLEC records and 0.004% of the BellSouth records.

Unadjusted
August BellSouth and CLEC Completion Interval-Provisioning
Non-Dispatched Cases





Descriptive Measures

Service Provider	Mean	Standard Deviation
BST	0.92	1.93
CLEC	1.41	1.94
Difference	-0.49	

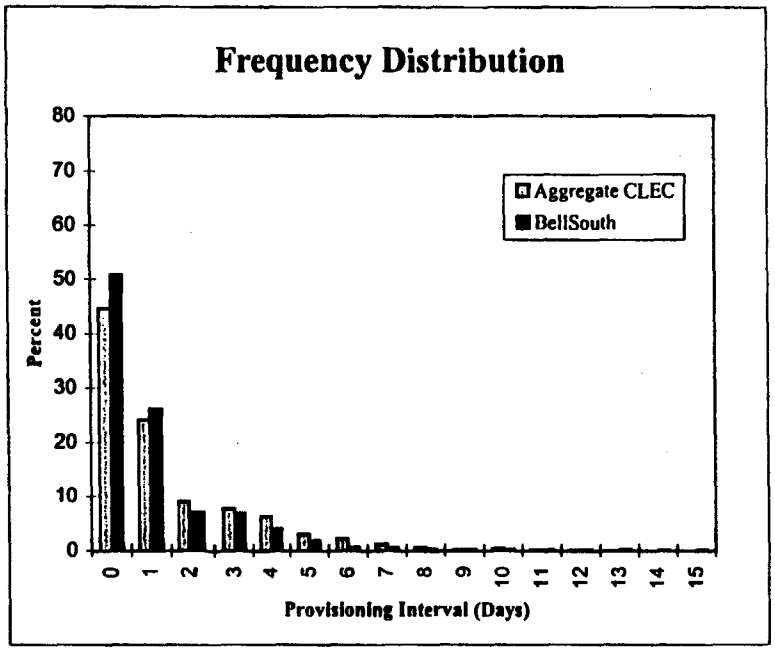
Analytic Measures

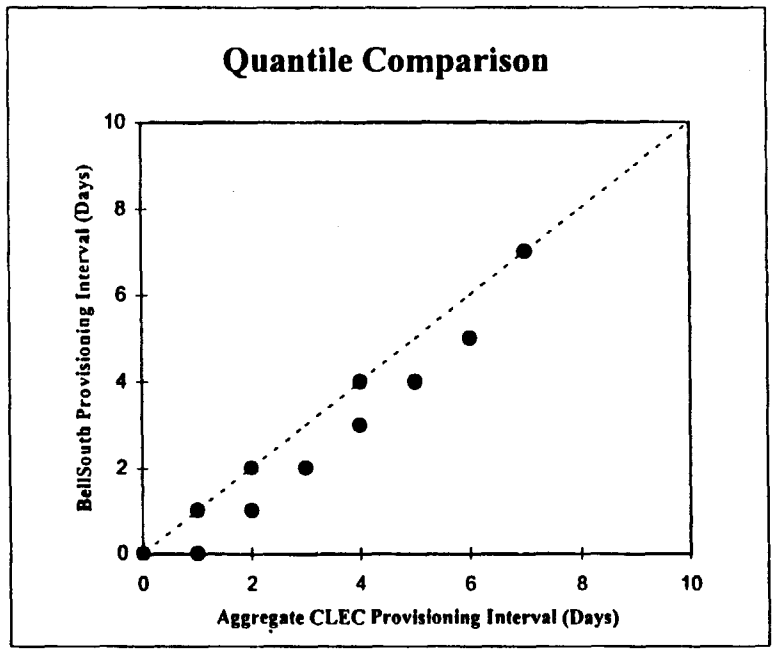
Testing Method	Test Statistic	P-value (percent)
LCUG	-30.41	0.0000
FCC	-30.41	0.0000
BST	-10.93	0.0000

Data used in analysis does not include any records with missed appointments due to customer rescheduling or records corresponding to official services.

The application of statistical trimming removed records with completion interval-provisioning of above 99 days. This resulted in the removal of no CLEC records and 0.004% of the BellSouth records.

Adjusted
August BellSouth and CLEC Completion Interval-Provisioning
Non-Dispatched Cases





Descriptive Measures

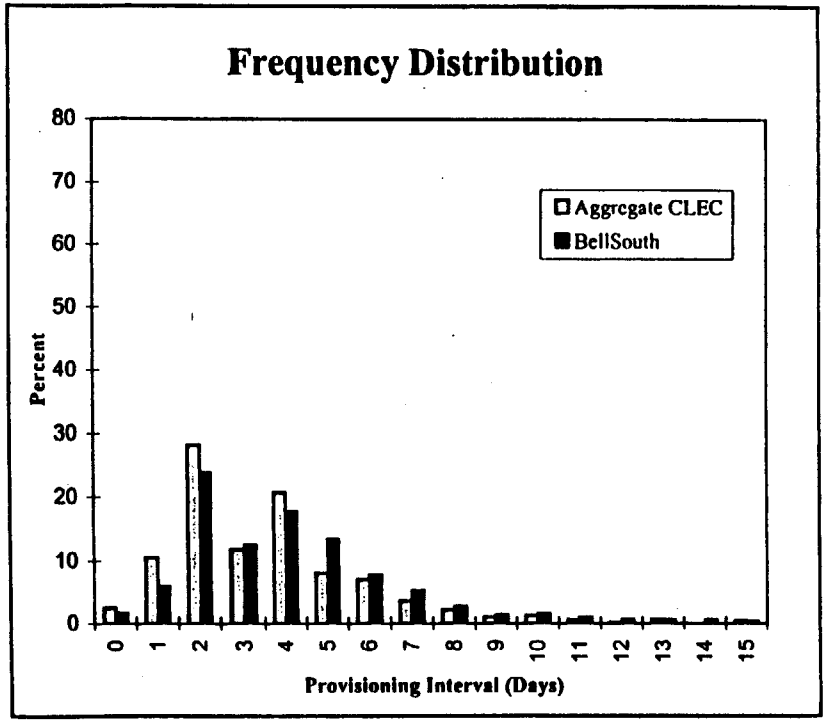
Service Provider	Mean	Standard Deviation
BST	1.18	2.33
CLEC	1.41	1.94
Difference	-0.23	

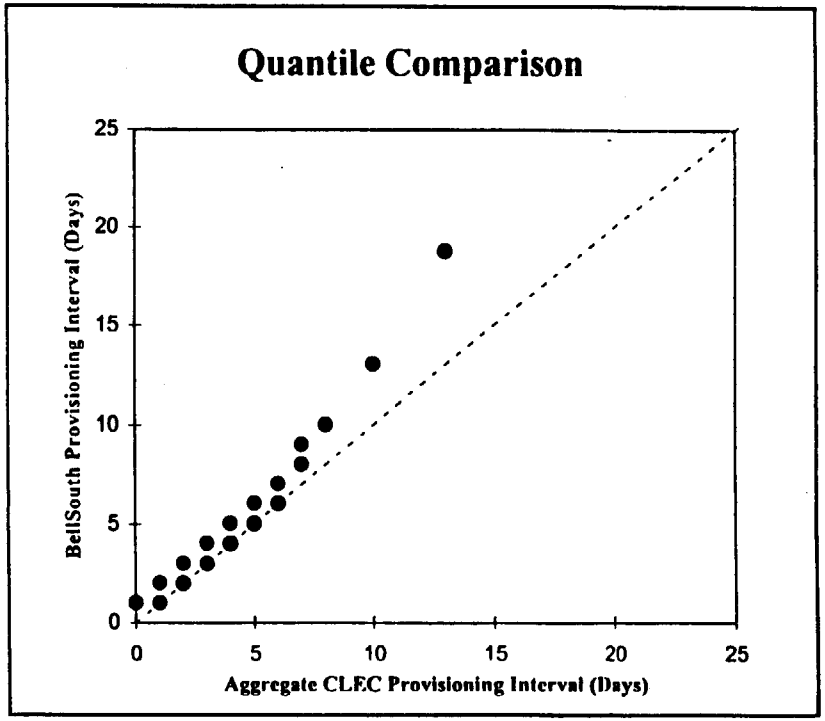
Analytic Measures

Testing Method	Test Statistic	P-value (percent)
LCUG	-11.86	0.0000
FCC	-11.93	0.0000
BST	-4.39	0.0068

Data used in analysis does not include any records with missed appointments due to customer rescheduling or records corresponding to official services.
The application of statistical trimming removed records with completion interval-provisioning of above 99 days. This resulted in the removal of no CLEC records and 0.004% of the BellSouth records.

Unadjusted
August BellSouth and CLEC Completion Interval-Provisioning
Dispatched, Residential, All Circuits





Descriptive Measures

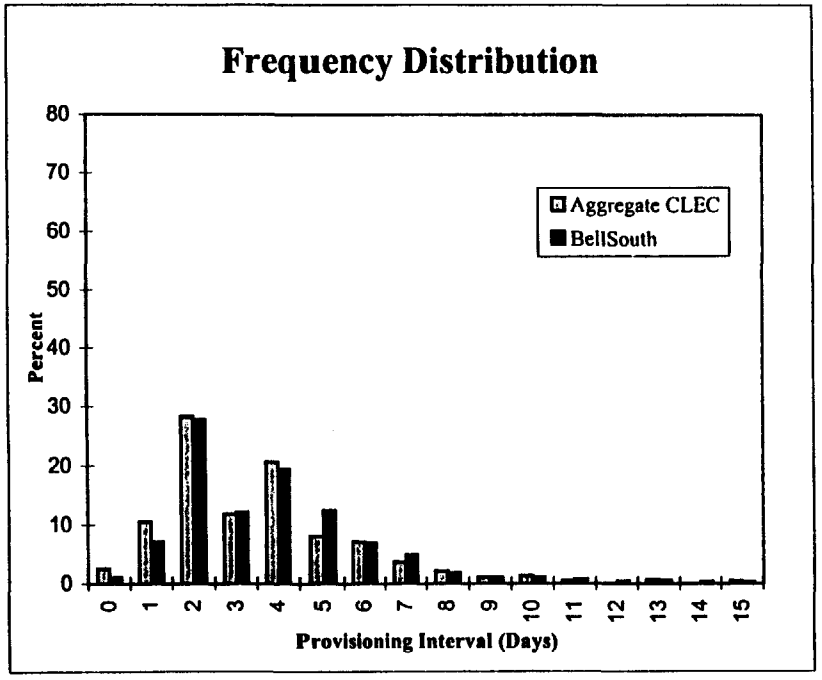
Service Provider	Mean	Standard Deviation
BST	4.69	4.45
CLEC	3.84	3.38
Difference	0.85	

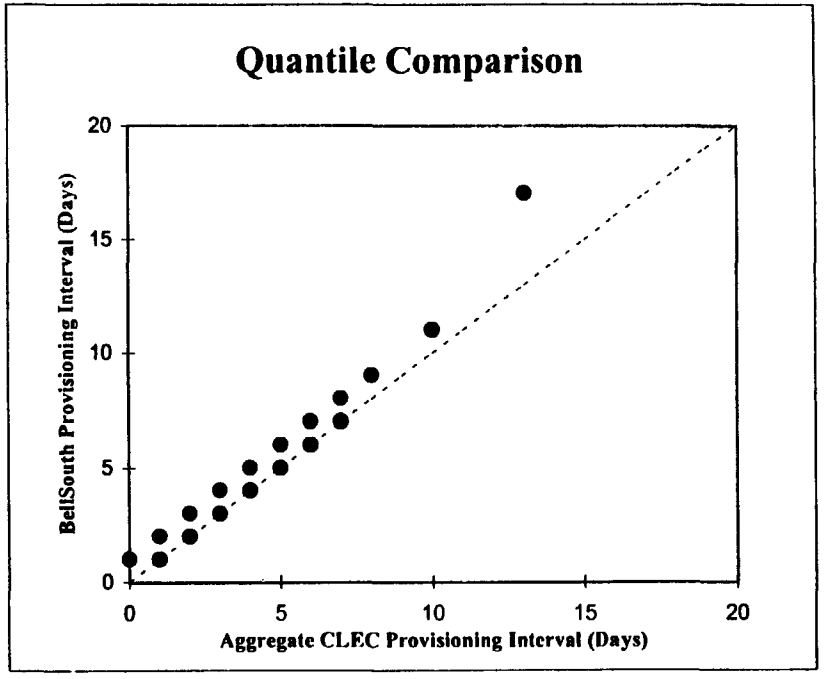
Analytic Measures

Testing Method	Test Statistic	P-value (percent)
LCUG	5.77	0.0000
FCC	5.83	0.0000
BST	8.67	0.0000

Data used in analysis does not include any records with missed appointments due to customer rescheduling or records corresponding to official services. The application of statistical trimming removed records with completion interval-provisioning of above 99 days. This resulted in the removal of no CLEC records and 0.004% of the BellSouth records.

Adjusted
August BellSouth and CLEC Completion Interval-Provisioning
Dispatched, Residential, All Circuits





Descriptive Measures

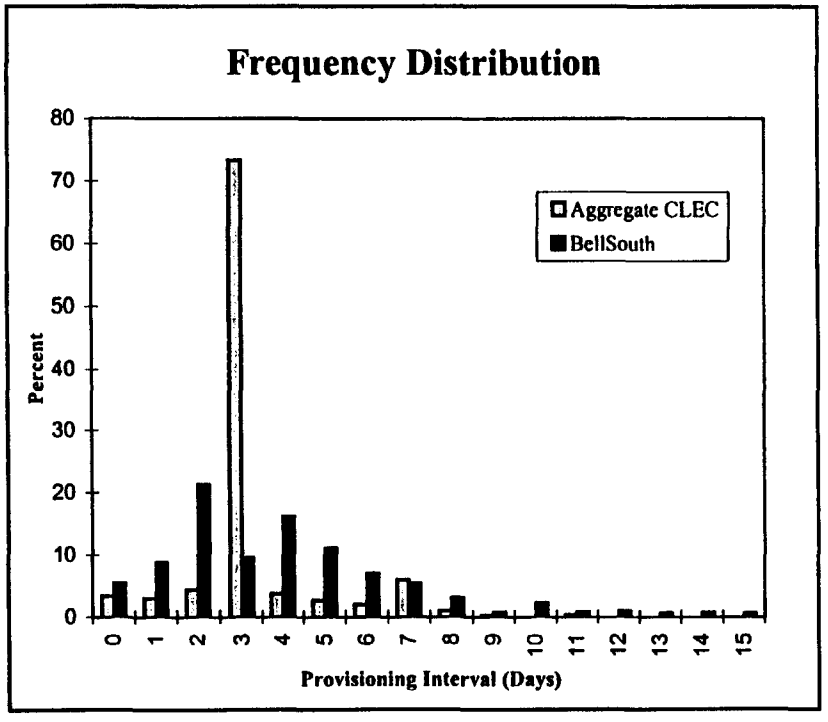
Service Provider	Mean	Standard Deviation
BST	4.34	4.19
CLEC	3.84	3.38
Difference	0.50	

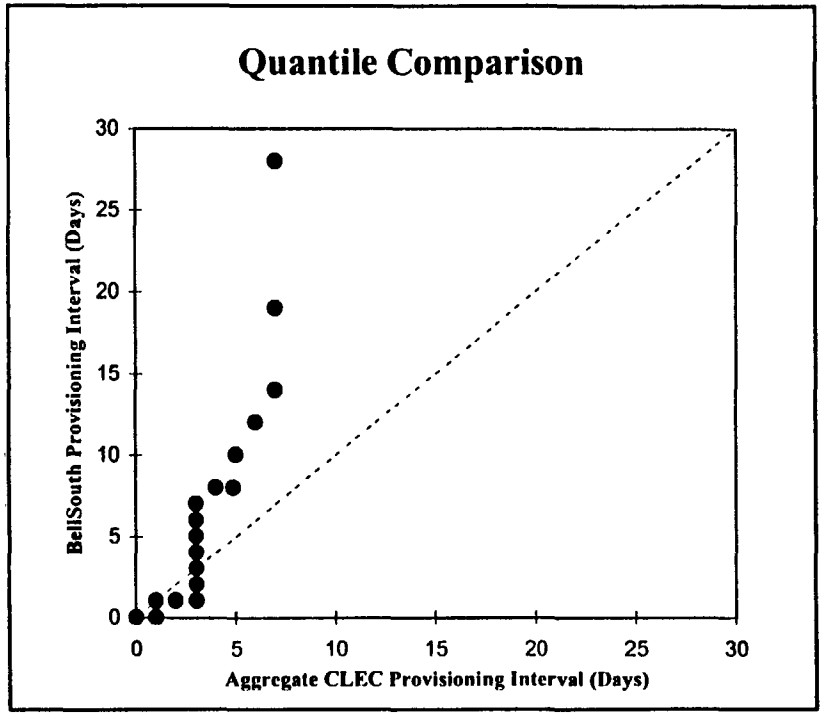
Analytic Measures

Testing Method	Test Statistic	P-value (percent)
LCUG	3.60	0.0159
FCC	3.63	0.0139
BST	4.40	0.0067

Data used in analysis does not include any records with missed appointments due to customer rescheduling or records corresponding to official services. The application of statistical trimming removed records with completion interval-provisioning of above 99 days. This resulted in the removal of no CLEC records and 0.004% of the BellSouth records.

Unadjusted
August BellSouth and CLEC Completion Interval-Provisioning
Dispatched, Business, All Circuits





Descriptive Measures

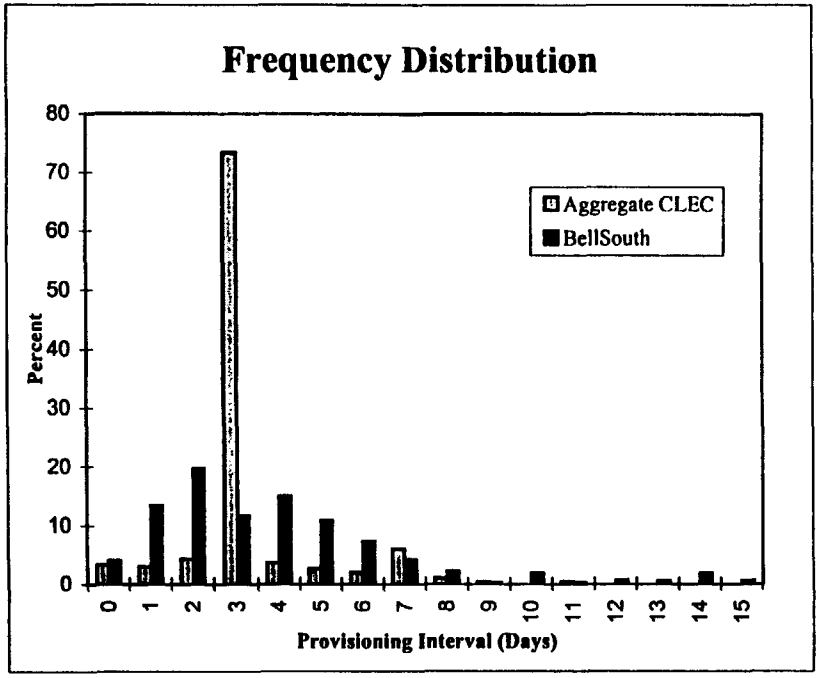
Service Provider	Mean	Standard Deviation
BST	5.37	7.34
CLEC	3.28	1.50
Difference	2.09	

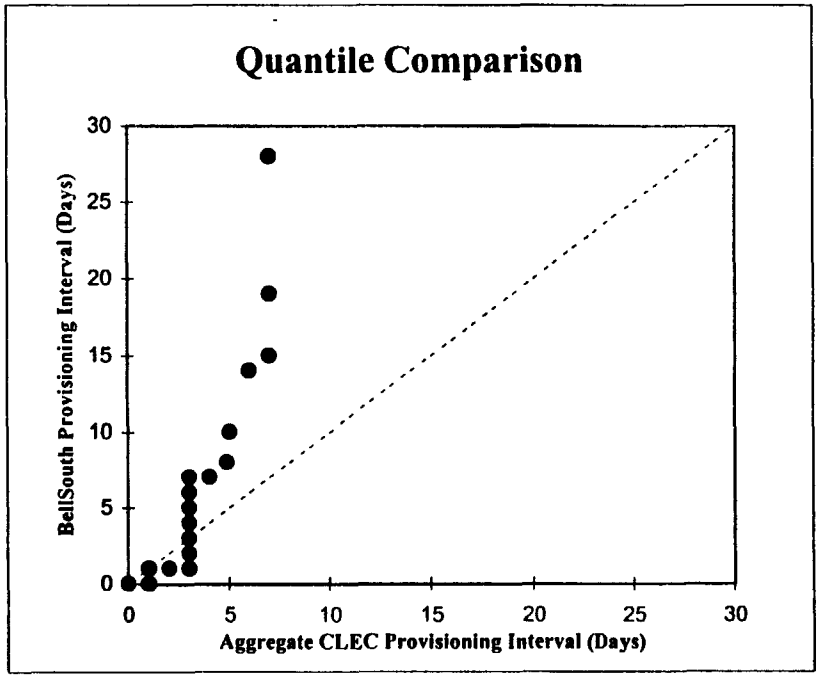
Analytic Measures

Testing Method	Test Statistic	P-value (percent)
LCUG	4.81	0.0001
FCC	4.93	0.0000
BST	8.86	0.0000

Data used in analysis does not include any records with missed appointments due to customer rescheduling or records corresponding to official services. The application of statistical trimming removed records with completion interval-provisioning of above 99 days. This resulted in the removal of no CLEC records and 0.004% of the BellSouth records.

Adjusted
August BellSouth and CLEC Completion Interval-Provisioning
Dispatched, Business, All Circuits





Descriptive Measures

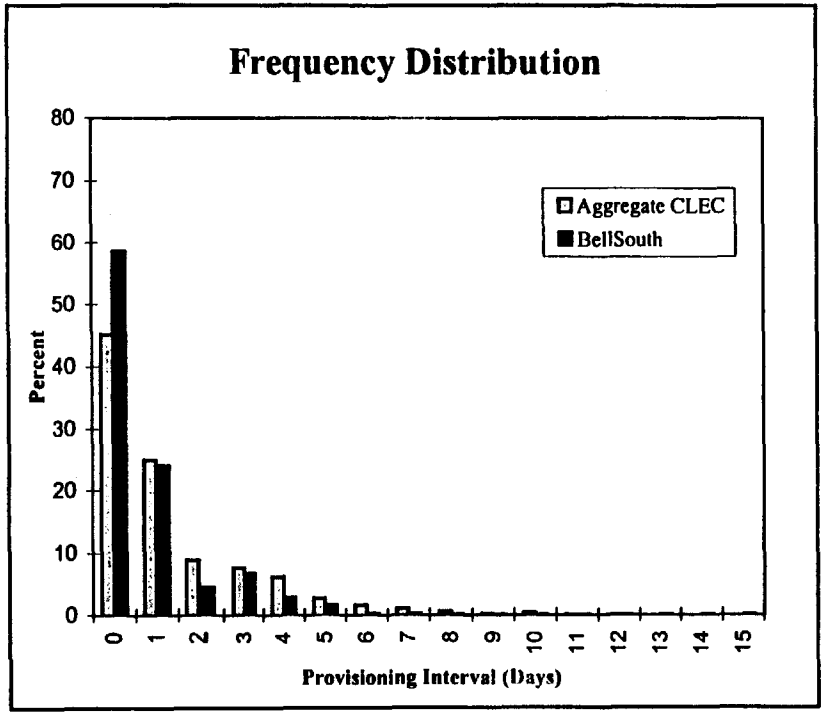
Service Provider	Mean	Standard Deviation
BST	5.27	7.25
CLEC	3.28	1.50
Difference	1.99	

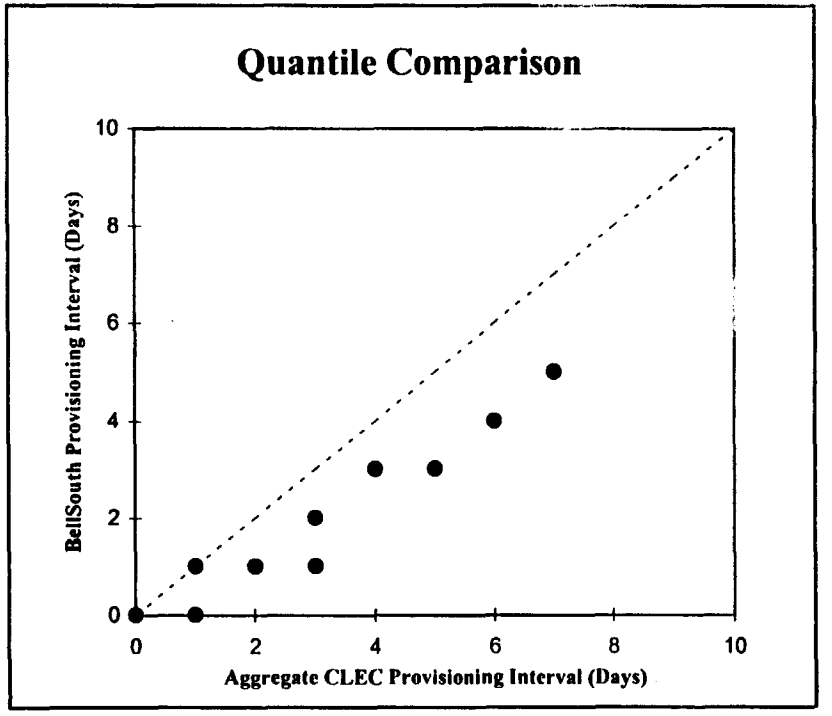
Analytic Measures

Testing Method	Test Statistic	P-value (percent)
LCUG	4.63	0.0002
FCC	4.75	0.0001
BST	2.48	0.9762

Data used in analysis does not include any records with missed appointments due to customer rescheduling or records corresponding to official services. The application of statistical trimming removed records with completion interval-provisioning of above 99 days. This resulted in the removal of no CLEC records and 0.004% of the BellSouth records.

Unadjusted
August BellSouth and CLEC Completion Interval-Provisioning
Non-Dispatched, Residential, All Circuits





Descriptive Measures

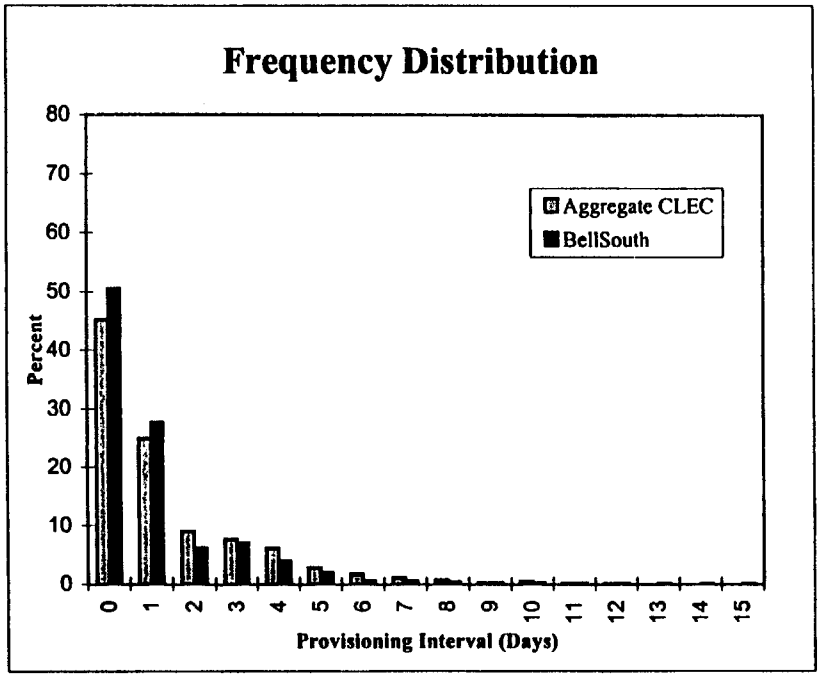
Service Provider	Mean	Standard Deviation
BST	0.88	1.83
CLEC	1.35	1.87
Difference	-0.47	

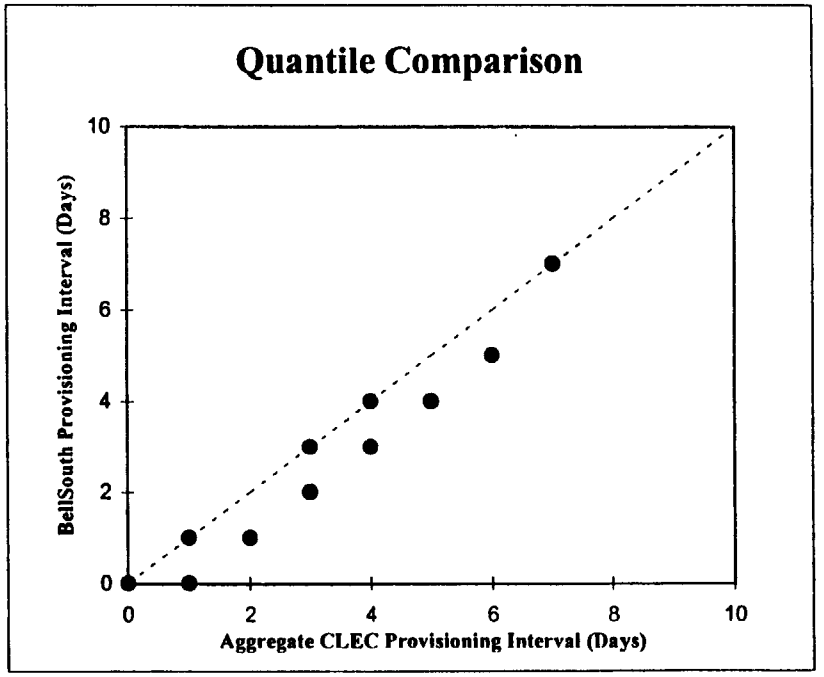
Analytic Measures

Testing Method	Test Statistic	P-value (percent)
LCUG	-29.48	0.0000
FCC	-29.46	0.0000
BST	-10.05	0.0000

Data used in analysis does not include any records with missed appointments due to customer rescheduling or records corresponding to official services.
The application of statistical trimming removed records with completion interval-provisioning of above 99 days. This resulted in the removal of no CLEC records and 0.004% of the BellSouth records.

Adjusted
August BellSouth and CLEC Completion Interval-Provisioning
Non-Dispatched, Residential, All Circuits





Descriptive Measures

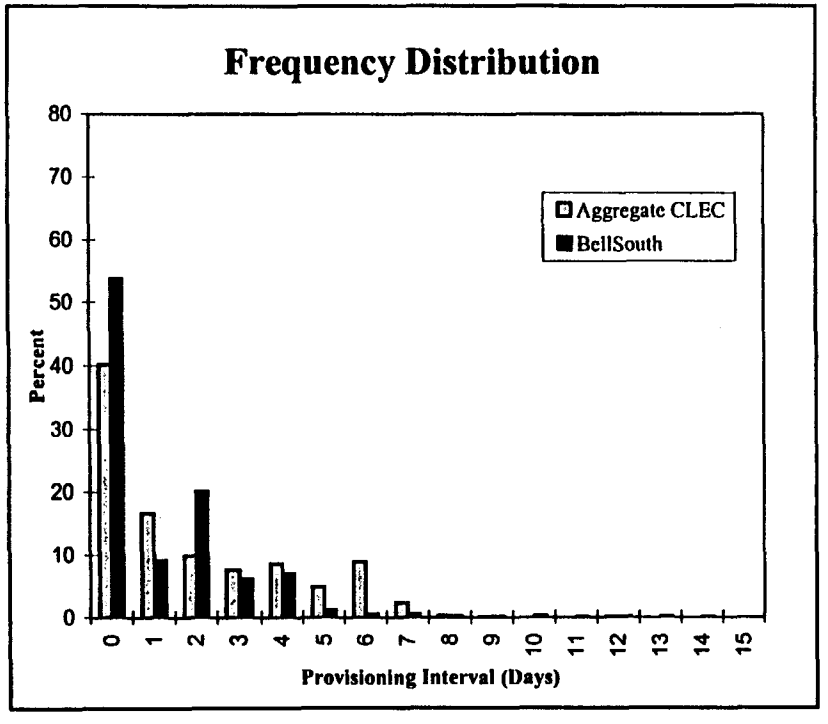
Service Provider	Mean	Standard Deviation
BST	1.15	2.26
CLEC	1.35	1.87
Difference	-0.20	

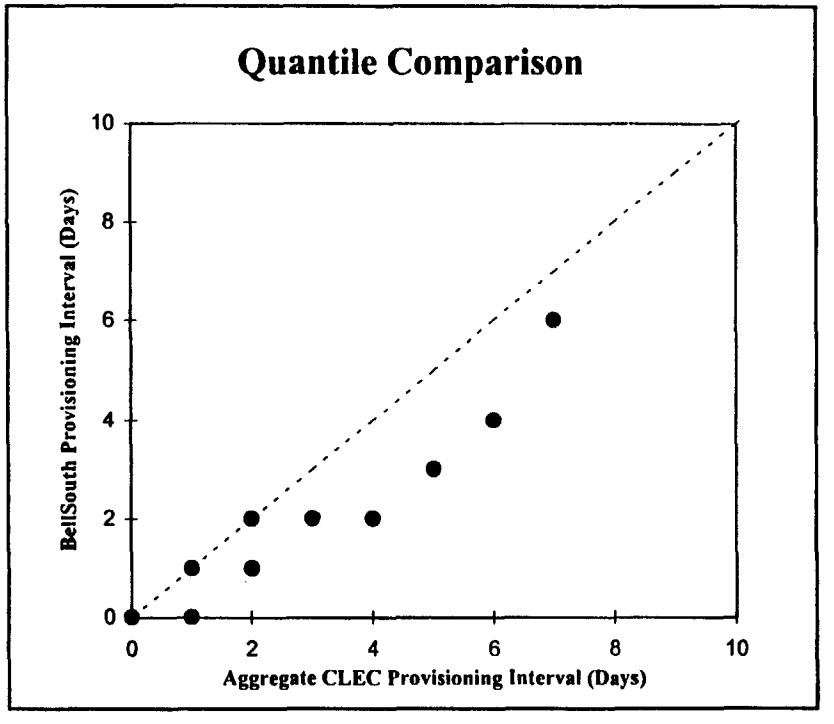
Analytic Measures

Testing Method	Test Statistic	P-value (percent)
LCUG	-10.38	0.0000
FCC	-10.44	0.0000
BST	-4.41	0.0066

Data used in analysis does not include any records with missed appointments due to customer rescheduling or records corresponding to official services. The application of statistical trimming removed records with completion interval-provisioning of above 99 days. This resulted in the removal of no CLEC records and 0.004% of the BellSouth records.

Unadjusted
August BellSouth and CLEC Completion Interval-Provisioning
Non-Dispatched, Business, All Circuits





Descriptive Measures

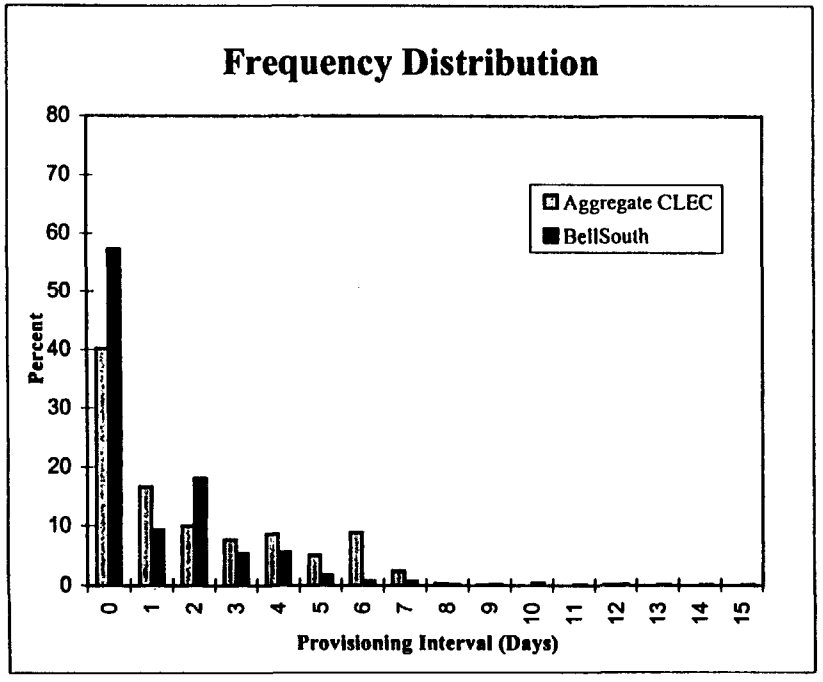
Service Provider	Mean	Standard Deviation
BST	1.28	2.65
CLEC	1.98	2.37
Difference	-0.70	

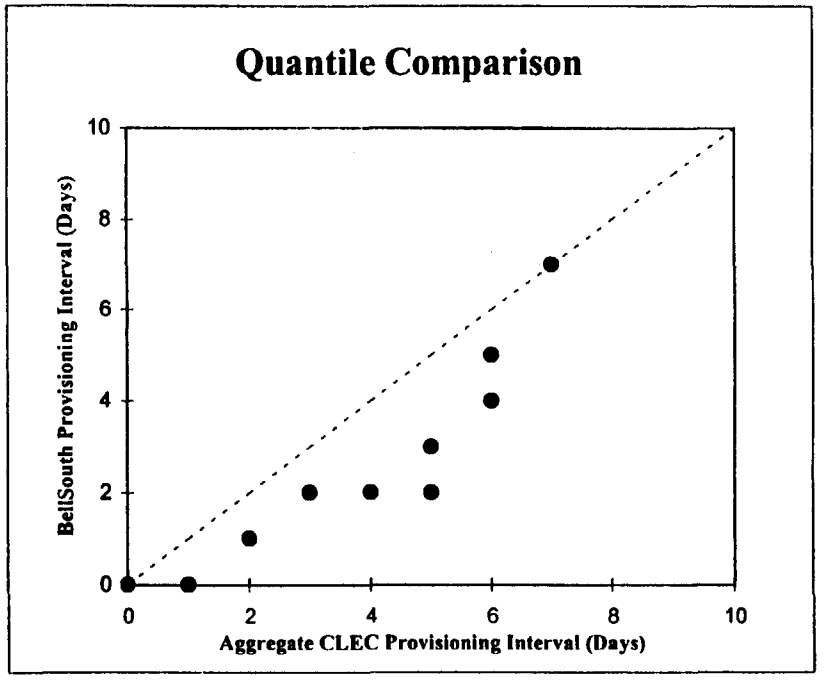
Analytic Measures

Testing Method	Test Statistic	P-value (percent)
LCUG	-8.69	0.0000
FCC	-8.72	0.0000
BST	-3.12	0.2098

Data used in analysis does not include any records with missed appointments due to customer rescheduling or records corresponding to official services. The application of statistical trimming removed records with completion interval-provisioning of above 99 days. This resulted in the removal of no CLEC records and 0.004% of the BellSouth records.

Adjusted
August BellSouth and CLEC Completion Interval-Provisioning
Non-Dispatched, Business, All Circuits





Descriptive Measures

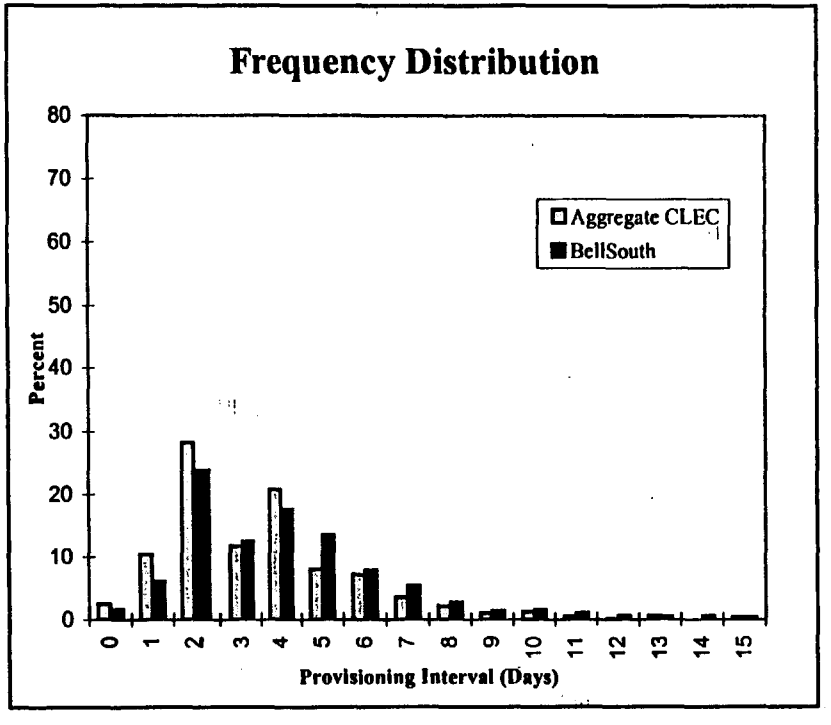
Service Provider	Mean	Standard Deviation
BST	1.20	2.47
CLEC	1.98	2.37
Difference	-0.78	

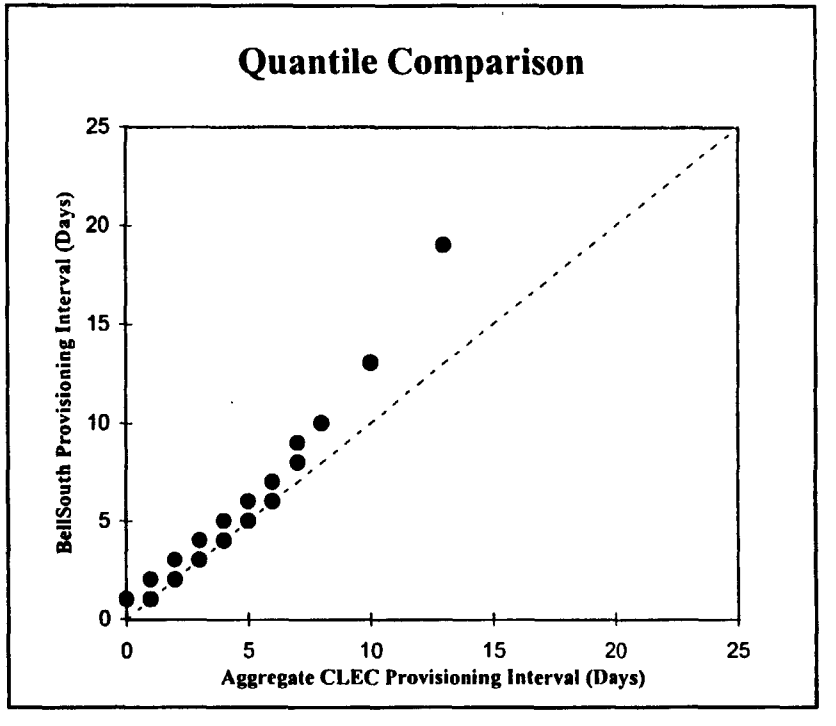
Analytic Measures

Testing Method	Test Statistic	P-value (percent)
LCUG	-10.42	0.0000
FCC	-10.43	0.0000
BST	-3.55	0.0686

Data used in analysis does not include any records with missed appointments due to customer rescheduling or records corresponding to official services. The application of statistical trimming removed records with completion interval-provisioning of above 99 days. This resulted in the removal of no CLEC records and 0.004% of the BellSouth records.

Unadjusted
August BellSouth and CLEC Completion Interval-Provisioning
Dispatched, Residential, Less Than 10 Circuits





Descriptive Measures

Service Provider	Mean	Standard Deviation
BST	4.70	4.45
CLEC	3.85	3.39
Difference	0.85	

Analytic Measures

Testing Method	Test Statistic	P-value (percent)
LCUG	5.73	0.0000
FCC	5.79	0.0000
BST	8.69	0.0000

Data used in analysis does not include any records with missed appointments due to customer rescheduling or records corresponding to official services. The application of statistical trimming removed records with completion interval-provisioning of above 99 days. This resulted in the removal of no CLEC records and 0.004% of the BellSouth records.